

Electrostatically Embedded Many-Body Expansion for Neutral and Charged Metalloenzyme Model Systems

Elbek K. Kurbanov, Hannah R. Leverentz, Donald G. Truhlar, and Elizabeth A. Amin*

Department of Medicinal Chemistry, Department of Chemistry, and Minnesota Supercomputing Institute, University of Minnesota, Minneapolis, Minnesota 55414, United States

S Supporting Information

ABSTRACT: The electrostatically embedded many-body (EE-MB) method has proven accurate for calculating cohesive and conformational energies in clusters, and it has recently been extended to obtain bond dissociation energies for metal–ligand bonds in positively charged inorganic coordination complexes. In the present paper, we present four key guidelines that maximize the accuracy and efficiency of EE-MB calculations for metal centers. Then, following these guidelines, we show that the EE-MB method can also perform well for bond dissociation energies in a variety of neutral and negatively charged inorganic coordination systems representing metalloenzyme active sites, including a model of the catalytic site of the zinc-bearing anthrax toxin lethal factor, a popular target for drug development. In particular, we find that the electrostatically embedded three-body (EE-3B) method is able to reproduce conventionally calculated bond-breaking energies in a series of pentacoordinate and hexacoordinate zinc-containing systems with an average absolute error (averaged over 25 cases) of only 0.98 kcal/mol.

Zinc is an essential transition metal required for the catalytic and structural activity of many enzymes,¹ and it participates in a number of key biological processes in living systems, including immune function,^{2,3} protein synthesis,^{2,5} wound healing,^{4,7} DNA synthesis,^{2,6} and cell division.^{2,6} Zinc metalloenzymes carry out essential functions in a wide variety of biochemical pathways and have attracted much attention as drug design targets; examples include the anthrax toxin lethal factor,⁸ insulin, phosphotriesterase, the matrix metalloproteinases, cytidine deaminase, histone deacetylases, zinc-finger proteins, and human carbonic anhydrase. In these enzymes, zinc may play structural and/or catalytic roles, with catalysis taking place in the first coordination shell.⁹ *In silico* techniques have generally proven valuable for rational drug design and enzyme modeling; however, reliable representation of zinc and other transition metal centers in macromolecules is nontrivial due to the complexity of the coordination environment and charge distribution at the catalytic center. Accurate zinc modeling requires quantum mechanical electronic structure calculations that pose challenges due to system size and the complexity of the calculations. Enabling accurate simulations on large and computationally demanding systems such as biozinc metallic coordination sites is a central focus of quantum chemistry research, and attempts have been made^{9–11} to assess the accuracy of various QM-based strategies for Zn model systems representing biocenters and other complex environments such as nanoparticles and clusters. Fragmentation is a useful strategy for addressing these roadblocks, and various schemes have been explored in order to reduce calculation complexity.^{12–24} The electrostatically embedded many-body expansion (EE-MB) method has emerged as a particularly promising approach.^{10,17,25–29} As described in our previous work,^{10,17,25–30} EE-MB addresses the challenge of system size by partitioning larger complexes into a series of fragments, embedding fragment energies in a field of point charges and running calculations in parallel.

In the EE-MB method,^{10,17,25–30} the fragments into which a system is partitioned are called monomers. In the present study, we examine two variants of this method: the electrostatically embedded pairwise additive (EE-PA) approximation and the electrostatically embedded three-body (EE-3B) approximation. In the former, the energy of systems composed of monomers m , n , p , ... is approximated as

$$E^{\text{PA}} = E^{(1)} + \Delta E^{(2)} \quad (1)$$

where

$$E^{(1)} = \sum_m E_m \quad (2)$$

$$\Delta E^{(2)} = \sum_m \sum_{n > m} \Delta E_{mn}^{(2)} \quad (3)$$

$$\Delta E_{mn}^{(2)} = E_{mn} - E_m - E_n \quad (4)$$

whereas the EE-3B energy is defined as

$$E^{3\text{B}} = E^{\text{PA}} + \Delta E^{(3)} \quad (5)$$

where

$$\Delta E^{(3)} = \sum_m \sum_{n > m} \sum_{p > n} \Delta_{mnp} \quad (6)$$

$$\Delta E_{mnp} = E_{mnp} - E_{mnp}^{\text{PA}} \quad (7)$$

where E_m , E_{mn} , and E_{mnp} are the energies of a monomer, dimer, and trimer, respectively, embedded in a field of point charges

Received: September 12, 2011

Published: November 29, 2011

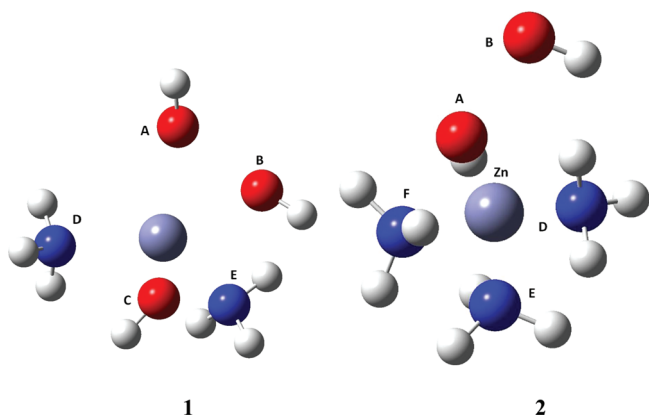


Figure 1. Structures of truncated model Zn biocenter complexes: (1) the anthrax toxin lethal factor active site (LF; 1PWU.pdb),²⁶ $[\text{Zn}(\text{NH}_3)_2(\text{OH})_3]^-$, and (2) matrix metalloproteinase-3 (MMP-3, stromelysin-1; 1SLN.pdb),²⁷ $[\text{Zn}(\text{NH}_3)_3(\text{OH})_2]$. In both cases, histidine residues are represented by ammonias (NH_3), and Glu residues and zinc-binding group (ZBG) oxygens in the cocrystallized inhibitors are represented by hydroxyls (OH^-).

representing the other monomers, and the individual energies are obtained using any type of electronic structure theory.

We have already shown that the EE-MB method can be used to calculate usefully accurate bond dissociation energies at low computational cost for positively charged Zn^{2+} systems; in particular the EE-3B method predicts bond energies obtained by conventional full-system calculations done at the same level of theory to within 1.0 kcal/mol for those cationic Zn^{2+} complexes.³⁰ In the present work, we recommend a set of specific fragmentation strategies to enhance the accuracy of EE-MB for coordination chemistry, and we assess the suitability of the EE-3B method for the more challenging neutral and negatively charged penta- and hexacoordinate Zn systems of biological importance. We also present EE-PA results for comparison.

Charges are calculated for each fragment at the geometry of that monomer in the overall system. For example, if we are calculating the energy of ZnABCDEF , where A, B, C, D, E, and F are ligands, and if one of the fragments is ZnBC , we calculate the partial atomic charges of ZnBC by removing A, D, E, and F from the system. Here, we calculate charges using Merz–Kollman (MK) electrostatic fitting,³⁷ as in previous work on Zn compounds.³⁰

All calculations were done with the M05-2X density functional³¹ and the B2 basis set,⁹ which is a polarized valence-triple- ζ basis set optimized and validated for use with Zn-containing complexes including biozinc coordination systems. Our earlier published work on a variety of Zn–ligand systems of importance in biology, nanotechnology, and drug design^{9,34} showed that incorporating relativistic effects on core electrons significantly increased the accuracy of geometric and energetic calculations for Zn coordination complexes; in the current study, we therefore replaced the 10 innermost electrons of Zn with the (MEFIT, R) relativistic effective core potential.^{32,33} The M05-2X/B2 density functional/basis set combination was chosen because of previous evaluations^{9,34} that yielded very accurate results for zinc complexes. We note explicitly, however, that the main objective of using DFT in this study is to assess whether the EE-MB approximation can reproduce full (unfragmented) calculations. If so, one could, for example, use the EE-MB approximation with

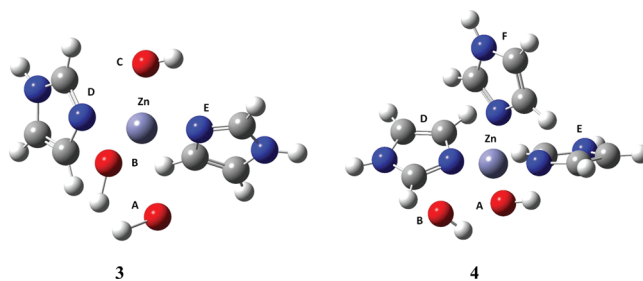


Figure 2. Structures of extended Zn biocenter complexes: (3) the anthrax toxin lethal factor active site (LF; 1PWU.pdb),²⁶ $[\text{Zn}(\text{Imd})_2(\text{OH})_3]^-$, and (4) matrix metalloproteinase-3 (MMP-3, stromelysin-1; 1SLN.pdb),²⁷ $[\text{Zn}(\text{Imd})_3(\text{OH})_2]$. In both cases, histidine residues are represented by imidazoles (Imd), and Glu residues and zinc-binding group (ZBG) oxygens in the cocrystallized inhibitors are represented by hydroxyls (OH^-).

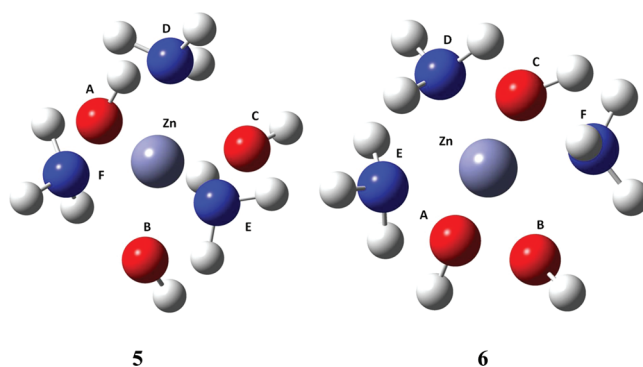


Figure 3. Structures of two octahedral, hexacoordinate Zn complexes ($[\text{Zn}(\text{NH}_3)_3(\text{OH})_3]^-$): (5) *fac* isomer and (6) *mer* isomer.

coupled cluster calculations on the fragments to approximate full coupled cluster calculations that are currently unaffordable.

All unfragmented calculations were performed using *Gaussian 09*.³⁵ All EE-MB calculations were carried out using MBPAC 2011–2,³⁶ an in-house software package that allows the user to define a particular fragmentation scheme and then accesses a locally modified version of *Gaussian 09* to perform the necessary monomer, dimer, and trimer calculations.

In the current work, we consider four pentacoordinate and two hexacoordinate Zn systems. The pentacoordinate complexes are model compounds based on experimental X-ray structures of two Zn metalloenzyme active sites relevant to biology and to the drug design process: the anthrax toxin lethal factor (LF; PDB ID: 1PWU)³⁸ and the matrix metalloproteinase-3 (MMP-3) catalytic site (PDB ID: 1SLN).³⁹ In LF, the catalytic Zn is coordinated by two histidines and one glutamic acid, and in 1PWU, the zinc is also ligated by two oxygens in the hydroxamate zinc-binding group (ZBG) of the cocrystallized inhibitor, forming the complete pentacoordinate system. In MMP-3, the catalytic zinc is similarly coordinated by three histidine residues, and in 1SLN, the two remaining coordination sites are occupied by the carboxylate ZBG of the cocrystallized inhibitor. We specifically chose pentacoordinate systems that include ligands from potential drug scaffold ZBGs, in order to test the ability of EE-MB to reproduce bond dissociation energies that would parallel the interactions of small molecules with drug-target catalytic centers.

We created two simple and two extended models of each biocenter, where the simple models **1** and **2** (Figure 1) represent His residues by ammonias and Glu side chains and ZBG oxygens by hydroxyls, yielding $[\text{Zn}(\text{NH}_3)_2(\text{OH})_3]^-$ as a model for the anthrax toxin lethal factor and $[\text{Zn}(\text{NH}_3)_3(\text{OH})_2]$ as a model for MMP-3. In the extended models **3** and **4** (Figure 2), the ammonias are replaced by full imidazole moieties while the hydroxyls are retained. The hexacoordinate complexes examined here are the *fac* and *mer* isomers of $[\text{Zn}(\text{NH}_3)_3(\text{OH})_3]^-$ (systems **5** and **6**, respectively, Figure 3). In total, these systems comprise four negatively charged and two neutral complexes. For systems **1**–**4**, all Zn–ligand distances were fixed at their experimental X-ray values. The hydrogen atoms on all NH_3 and OH ligands, and all Zn–ligand distances in systems **5** and **6**, were placed at standard distances and default orientations by the *GaussView*⁴⁰ program. The default N–H bond length in NH_3 is 1.00 Å. The default O–H distance is 0.96 Å. For systems **5** and **6**, the default Zn– NH_3 bond length is 1.95 Å, and the default Zn–OH distance is 1.91 Å. Default bond angles for ligand geometries in *GaussView* are obtained through AM1 optimizations. All structures are provided in the Supporting Information.

The quantity we calculate is a relative bond dissociation energy, which is defined as the energy to remove one of the ligands from the coordination system. As discussed in previous work,³⁰ this quantity is the sum of the energies of the two products (separated frozen fragments) minus the energy of the reactant, without including vibrational energy (thus, it is D_e , not D_0). When calculating the energies of a given dissociation product, the embedding charges of the other product are not included because the other product is considered to be infinitely separated.

After performing extensive calculations with various fragmentation schemes on systems **1** and **2**, we established four key fragmentation guidelines that, when applied, yielded the best results for all six systems in the current work. Next, we present these four guidelines.

First, our calculations on neutral and negatively charged Zn systems demonstrate, consistently with our previous findings,³⁰ that one must choose a fragmentation scheme where one of the monomers is Zn^{2+} coordinated to at least two ligands. We rationalize this rule in terms of partial atomic charges. In particular, the charge on unligated or monoligated Zn and even on biligated Zn is much larger than the charge on polyligated Zn; thus fragments consisting of unligated, monoligated, or—to a lesser extent—biligated Zn would not be representative of a portion of a larger system. But if each fragment already has two ligands on Zn, then even in dimers there are three ligands on Zn.

Second, as a corollary to rule 1, we do not dissociate bonds within fragments, as that would result in a product with Zn connected to a single ligand.

Third, at most, one fragment can be charged. We rationalize rule 3 as eliminating the longest-range electrostatic effects.

Finally, our fourth guideline allows no *trans* coordination; i.e., Zn^{2+} cannot be coordinated within a fragment with two ligands that are *trans* to each other. This rule can be understood as requiring links to be compact, although its origin is purely empirical at present.

We use the labeling scheme defined by Figures 1–3, in which A, B, and C (when present) are negatively charged hydroxyl ligands and D, E, and F (when present) are neutral ligands. A consequence of rule 3 for the present study is that Zn^{2+} coupled

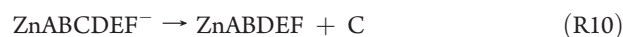
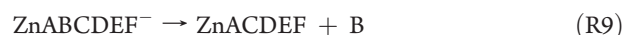
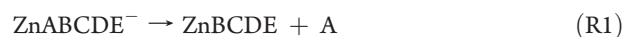
Table 1. Systems Considered in This Work and the Largest Fragment in Each^a

full system	largest fragment
$[\text{Zn}(\text{NH}_3)_2(\text{OH})_3]^-$	$\text{Zn}(\text{OH})_2$ (ZnBC)
$[\text{Zn}(\text{NH}_3)_2(\text{OH})_3]^-$	$\text{Zn}(\text{OH})_2$ (ZnAB)
$[\text{Zn}(\text{NH}_3)_2(\text{OH})_3]^-$	$\text{Zn}(\text{OH})_2$ (ZnBC, ZnAB)
$[\text{Zn}(\text{NH}_3)_3(\text{OH})_2]$	$\text{Zn}(\text{OH})_2$ (ZnAB)
$[\text{Zn}(\text{Imd})_2(\text{OH})_3]^-$	$\text{Zn}(\text{OH})_2$ (ZnBC)
$[\text{Zn}(\text{Imd})_2(\text{OH})_3]^-$	$\text{Zn}(\text{OH})_2$ (ZnAB)
$[\text{Zn}(\text{Imd})_2(\text{OH})_3]^-$	$\text{Zn}(\text{OH})_2$ (ZnBC, ZnAB)
$[\text{Zn}(\text{Imd})_3(\text{OH})_2]$	$\text{Zn}(\text{OH})_2$ (ZnAB)
<i>fac</i> isomer of $[\text{Zn}(\text{NH}_3)_3(\text{OH})_3]^-$	$\text{Zn}(\text{OH})_2$ (ZnBC)
<i>fac</i> isomer of $[\text{Zn}(\text{NH}_3)_3(\text{OH})_3]^-$	$\text{Zn}(\text{OH})_2$ (ZnAC)
<i>fac</i> isomer of $[\text{Zn}(\text{NH}_3)_3(\text{OH})_3]^-$	$\text{Zn}(\text{OH})_2$ (ZnAB)
<i>fac</i> isomer of $[\text{Zn}(\text{NH}_3)_3(\text{OH})_3]^-$	$\text{Zn}(\text{OH})_2$ (ZnAB, ZnBC, ZnAC)
<i>mer</i> isomer of $[\text{Zn}(\text{NH}_3)_3(\text{OH})_3]^-$	$\text{Zn}(\text{OH})_2$ (ZnBC)
<i>mer</i> isomer of $[\text{Zn}(\text{NH}_3)_3(\text{OH})_3]^-$	$\text{Zn}(\text{OH})_2$ (ZnAB)
<i>mer</i> isomer of $[\text{Zn}(\text{NH}_3)_3(\text{OH})_3]^-$	$\text{Zn}(\text{OH})_2$ (ZnAB, ZnBC)

^aWhen there is more than one row for a given system, it is because the largest fragment is not the same in all calculations on that system.

with two hydroxyl groups must be part of the fragmentation scheme in all six complexes.

Rules 3 and 4, taken together, forbid applying EE-MB to the dissociation of monomer B in **1** or monomer B in **6** because rule 3 would then require Zn to be coordinated within a fragment to ligands A and C in **1** and to ligands A and C in **6**, which in both cases would violate rule 4. After eliminating these processes that cannot be treated by the guidelines, we consider all of the remaining processes, which may be classified as follows:



Keeping the four guidelines in mind, we considered dissociation processes R1–R4 for systems **1** and **3**, processes R5–R7 for

Table 2. Benchmark Bond Energies (kcal/mol)

reaction	system	dissociated bond	bond energy	largest Zn fragment(s) in rxn
R1	1	Zn–A	35.12	ZnBC
R2	1	Zn–C	70.22	ZnAB
R3	1	Zn–D	19.32	ZnBC, ZnAB
R4	1	Zn–E	15.89	ZnBC, ZnAB
R5	2	Zn–D	–5.28	ZnAB
R6	2	Zn–E	–13.49	ZnAB
R7	2	Zn–F	6.19	ZnAB
R1	3	Zn–A	12.03	ZnBC
R2	3	Zn–C	57.01	ZnAB
R3	3	Zn–D	17.51	ZnBC, ZnAB
R4	3	Zn–E	20.68	ZnBC, ZnAB
R5	4	Zn–D	20.3	ZnAB
R6	4	Zn–E	–7.26	ZnAB
R7	4	Zn–F	9.53	ZnAB
R8	5	Zn–A	10.47	ZnBC
R9	5	Zn–B	9.18	ZnAC
R10	5	Zn–C	10.47	ZnAB
R11	5	Zn–D	–15.25	ZnAB, ZnBC, ZnAC
R12	5	Zn–E	–22.28	ZnAB, ZnBC, ZnAC
R13	5	Zn–F	–20.74	ZnAB, ZnBC, ZnAC
R8	6	Zn–A	26.98	ZnBC
R10	6	Zn–C	34.78	ZnAB
R11	6	Zn–D	–15.65	ZnAB, ZnBC
R12	6	Zn–E	–17.64	ZnAB, ZnBC
R13	6	Zn–F	–10.05	ZnAB, ZnBC

Table 3. Unsigned Errors in Bond Energies (kcal/mol) for Systems 1 and 3

	EE-PA		EE-3B	
	1	3	1	3
R1	6.47	4.60	0.85	0.85
R2	6.81	8.71	0.78	1.24
R3	4.62	1.67	0.82	1.01
R4	1.38	4.01	0.81	1.04
mean	4.82	4.75	0.82	1.03

systems 2 and 4, and processes R8–R13 for systems 5 and 6, except for system 6, where process R9 was not considered because it would result in a monomer with ligands A and C positioned *trans* to each other. All systems considered in this work, together with the largest fragment in each, are listed in Table 1.

Benchmark values for bond dissociation energies were obtained by full single-point calculations, i.e., without using the many-body approximation (see Table 2). Note that both the benchmark and the many-body calculations employ the same M05-2X/B2/MEFIT,R method. We measure “errors” as the deviation of the EE-MB results from the full calculations with the same method. If the error is small, then we assume that the method could be used with confidence for systems where full calculations on the entire system are impractically expensive or undoable, either due to system size (larger ligands, entire

Table 4. Unsigned Errors in Bond Energies (kcal/mol) for Systems 2 and 4

	EE-PA		EE-3B	
	2	4	2	4
R5	2.37	5.84	1.10	0.81
R6	5.41	5.57	1.09	0.85
R7	1.91	5.91	1.08	0.83
mean	3.23	5.77	1.09	0.83

Table 5. Unsigned Errors in Bond Energies (Kcal/mol) for Systems 5 and 6

	EE-PA		EE-3B	
	5	6	5	6
R8	8.80	6.57	0.59	1.38
R9	9.26		0.37	
R10	8.73	7.54	0.05	2.20
R11	4.51	3.06	1.44	0.16
R12	4.49	2.40	1.40	1.00
R13	4.32	2.53	1.54	1.30
mean ^a	6.68	4.42	0.90	1.21

^a Mean unsigned error for the five or six cases in the given column.

metalloenzymes) or due to using a higher level of electronic structure theory, for example, coupled cluster theory.

Tables 3, 4, and 5 show the EE-MB bond-breaking energies and mean unsigned errors for all six systems. The systems are quite different, but the performance of the EE-3B method is uniformly good. For example, for 1, the bond dissociation energies range from 16 to 70 kcal/mol, but the error of the EE-3B method is in the range 0.78–0.85 kcal/mol for all four cases. The EE-3B method has a mean unsigned error (MUE) in bond dissociation energy of 0.82 kcal/mol for system 1, 1.09 kcal/mol for system 2, 1.03 kcal/mol for system 3, and 0.83 kcal/mol for system 4. It is encouraging that the EE-3B method performs very well for both the “truncated” model systems 1 and 2 and the “extended” model systems 3 and 4. The MUEs in bond dissociation energies for the hexacoordinate systems 5 and 6 are comparable to those for the pentacoordinate systems, at 0.90 and 1.21 kcal/mol, respectively. As expected, the EE-PA method is less accurate, resulting in MUEs in bond dissociation energies ranging from 3.23 to 6.68 kcal/mol for the systems studied here. Altogether, there are 25 cases in Tables 3, 4, and 5, and averaging the unsigned errors over all 25 gives an overall mean unsigned error of 5.10 kcal/mol for the EE-PA method but only 0.98 kcal/mol for the EE-3B method.

The EE-3B method, when applied using our fragmentation guidelines, reliably yields bond dissociation energies within 1.21 kcal/mol of full-calculation DFT benchmark values, further demonstrating its utility and accuracy for neutral and negatively charged bioinorganic structures, in addition to the positively charged systems evaluated in our previous work. Moreover, EE-MB exhibits high accuracy for “extended” active site models with His residues represented by full imidazole rings rather than ammonias, and for hexacoordinate Zn complexes, indicating its particular usefulness for larger metalloprotein active site systems for which full, high-level electronic structure calculations might

be intractable or may incur a high computational cost. Finally, EE-MB is likely to find use in the drug discovery process; it performs very well for pentacoordinate systems representing a small-molecule drug lead coordinated to a catalytic metal center (which are otherwise quite challenging to model), and it can also be used to obtain key parameters such as bond dissociation energies that can be imported into molecular mechanics force fields to increase the accuracy of simpler and less costly calculations on macromolecular drug targets.

■ ASSOCIATED CONTENT

S Supporting Information. Cartesian coordinates for all Zn systems addressed in this study. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: eamin@umn.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENT

The authors express their appreciation to Bo Wang for helpful discussions. This work was supported in part by the National Institutes of Health (R01 AI083234 to E.A.A.), by the University of Minnesota Department of Medicinal Chemistry, and the University of Minnesota Supercomputing Institute for Advanced Computational Research. This work was also supported in part by NSF Grant No. CHE09-56776.

■ REFERENCES

- (1) Sandstead, H. H. *J. Lab. Clin. Med.* **1994**, *124*, 322–327.
- (2) Prasad, A. S. *Nutrition* **1995**, *11*, 93.
- (3) Solomons, N. W. *Nutr. Rev.* **1998**, *56*, 27–28.
- (4) Heyneman, C. A. *Ann. Pharmacotherapy* **1996**, *30*, 186–187.
- (5) MacDonald, R. S. *J. Nutr.* **2000**, *130*, 1500S–1508S.
- (6) Wallwork, J. C.; Duerre, J. A. *J. Nutr.* **1985**, *115*, 252–262.
- (7) Henkin, R. I. N. *Engl. J. Med.* **1974**, *291*, 675–676.
- (8) Chiu, T. L.; Solberg, J.; Patil, S.; Geders, T. W.; Zhang, X.; Rangarajan, S.; Francis, R.; Finzel, B. C.; Walters, M. A.; Hook, D. J.; Amin, E. A. *J. Chem. Inf. Model.* **2009**, *49*, 2726–2734.
- (9) Amin, E. A.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 75.
- (10) Sorkin, A.; Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 683.
- (11) Brothers, E. N.; Suarez, D.; Deerfield, D. W., II; Merz, K. M., Jr. *J. Comput. Chem.* **2004**, *25*, 1677.
- (12) Mayhall, N. J.; Raghavachari, K. *J. Chem. Theory Comput.* **2011**, *7*, 1336–1343.
- (13) Zhang, D. W.; Zhang, J. Z. H. *J. Chem. Phys.* **2003**, *119*, 3599.
- (14) Fedorov, D. G.; Kitaura, K. *J. Phys. Chem. A* **2004**, *120*, 6832.
- (15) Li, S.; Li, W.; Fang, T. *J. Am. Chem. Soc.* **2005**, *127*, 7215.
- (16) Bettens, R. P. A.; Lee, A. M. *J. Phys. Chem. A* **2006**, *110*, 8777.
- (17) Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 46.
- (18) Collins, M. A.; Deev, V. A. *J. Chem. Phys.* **2006**, *125*, 104104.
- (19) Fedorov, D. G.; Kitaura, K. *J. Phys. Chem. A* **2007**, *111*, 6904.
- (20) Hirata, S.; Yagi, K. *Chem. Phys. Lett.* **2008**, *464*, 123.
- (21) Xie, W.; Song, L.; Truhlar, D. G.; Gao, J. *J. Chem. Phys.* **2008**, *128*, 234108.
- (22) Gordon, M. S.; Mullin, J. M.; Pruitt, S. R.; Roskop, L. B.; Slipchenko, L. V.; Boatz, J. A. *J. Phys. Chem. B* **2009**, *113*, 9646.

- (23) Söderhjelm, P.; Aquilante, F.; Ryde, U. *J. Phys. Chem. B* **2009**, *113*, 11085.
- (24) Li, W.; Piecuch, P. *J. Phys. Chem. A* **2010**, *114*, 6721.
- (25) Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 1342.
- (26) Dahlke, E. E.; Leverentz, H. R.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 33.
- (27) Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 1.
- (28) Leverentz, H. R.; Truhlar, D. G. *J. Chem. Theory Comput.* **2009**, *5*, 1573.
- (29) Speetzen, E. D.; Leverentz, H. R.; Lin, H.; Truhlar, D. G. In *Accurate Condensed Phase Electronic Structure Theory*; Manby, F., Ed.; CRC Press: Boca Raton, FL, 2010.
- (30) Hua, D.; Leverentz, H. R.; Amin, E. A.; Truhlar, D. G. *J. Chem. Theory Comput.* **2011**, *7*, 251–255.
- (31) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215–241. Erratum: **2008**, *119*, 525.
- (32) Dolg, M.; Wedig, U.; Stoll, H.; Preuss, H. *J. Chem. Phys.* **1987**, *86*, 866.
- (33) Kaupp, M.; Stoll, H.; Preuss, H. *J. Comput. Chem.* **1990**, *11*, 1029.
- (34) Sorkin, A.; Amin, E. A.; Truhlar, D. G. *J. Chem. Theory Comput.* **2009**, *5*, 1254.
- (35) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, Revision A.01; Gaussian, Inc.: Wallingford, CT, 2009.
- (36) Dahlke, E. E.; Lin, H.; Leverentz, H.; Truhlar, D. G. *MBPAC 2011–2*; University of Minnesota: Minneapolis, MN, 2011.
- (37) Besler, B. H.; Merz, K. M., Jr.; Kollman, P. A. *J. Comput. Chem.* **1990**, *11*, 431.
- (38) Turk, B. E.; Wong, T. Y.; Schwarzenbacher, R.; Jarrell, E. T.; Leppla, S. H.; Collier, J.; Liddington, R. C.; Cantley, L. C. *Nat. Struct. Mol. Biol.* **2003**, *11*, 60–66.
- (39) Becker, J. W.; Marcy, A. I.; Rokosz, L. L.; Axel, M. G.; Burbaum, J. J.; Fitzgerald, P. M. D.; Cameron, P. M.; Esser, C. K.; Hagmann, W. K.; Hermes, J. D.; Springer, J. P. *Protein Sci.* **1995**, *4*, 1966–1976.
- (40) Dennington, R.; Keith, T.; Millam, J. Semicem Inc.: Shawnee Mission, KS, 2009.

■ NOTE ADDED AFTER ASAP PUBLICATION

This paper was published on the Web on December 2, 2011, with missing grant information. The corrected version was reposted on December 30, 2011.

Overcoming the Barrier on Time Step Size in Multiscale Molecular Dynamics Simulation of Molecular Liquids

Igor P. Omelyan^{*,†,‡,§} and Andriy Kovalenko^{*,‡,†}

[†]Department of Mechanical Engineering, University of Alberta, Mechanical Engineering Bldg. 4-9, Edmonton, AB, T6G 2G8, Canada

[‡]National Institute for Nanotechnology, 11421 Saskatchewan Dr., Edmonton, Alberta, T6G 2M9, Canada

[§]Institute for Condensed Matter Physics, National Academy of Sciences of Ukraine, 1 Svientsitskii Street, UA-79011, Lviv, Ukraine

ABSTRACT: We propose and validate a new multiscale technique, the extrapolative isokinetic Nosé–Hoover chain orientational (EINO) motion multiple time step algorithm for rigid interaction site models of molecular liquids. It nontrivially combines the multiple time step decomposition operator method with a specific extrapolation of intermolecular interactions, complemented by an extended isokinetic Nosé–Hoover chain approach in the presence of translational and orientational degrees of freedom. The EINO algorithm obviates the limitations on time step size in molecular dynamics simulations. While the best existing multistep algorithms can advance from a 5 fs single step to a maximum 100 fs outer step, we show on the basis of molecular dynamics simulations of the TIP4P water that our EINO technique overcomes this barrier. Specifically, we have achieved giant time steps on the order of 500 fs up to 5 ps, which now become available in the study of equilibrium and conformational properties of molecular liquids without a loss of stability and accuracy.

1. INTRODUCTION

The method of molecular dynamics (MD) is one of the most powerful tools for the investigation of various properties in liquids.^{1–4} This especially concerns such systems as water and its solutions as well as more complex biophysical fluids, including solvated proteins, which are of interest for modern chemistry and medicine. A characteristic feature of these systems is the coexistence of dynamical processes with vastly different time scales, extending from femtoseconds up to the millisecond region.^{5–8} For instance, in water, the fastest motion relates to the intramolecular vibrations of atoms, an intermediate scale arises from the strong short-range intermolecular potentials, while slow dynamics appears due to the weak long-range van der Waals and Coulombic interactions.

In MD simulations, however, the size of time steps is restricted to rather small values in order to avoid numerical instabilities and achieve a desired accuracy when integrating the equations of motion. This imposes severe limitations on the efficiency of MD calculations. It is obvious that longer time steps are more preferable because then the number of discretization points decreases, reducing the computational costs. Many multiple time stepping (MTS) techniques have been devised over the years to enlarge the time step size during the MD integration. They include the generalized Verlet integrator,⁹ reversible reference system propagator algorithm (RESPA),^{10–12} normal mode theories,^{13–16} mollified impulse schemes,^{17–19} Langevin dynamics approaches,^{20–22} and canonical Nosé–Hoover-like^{23–26} and isokinetic²⁷ thermostatting versions of RESPA, as well as more recent developments.^{28,29}

The MTS concept implies that faster components of motion are integrated with inner (smaller) time steps with respect to an outer (larger) one which is employed to handle slow dynamics. This leads to a speedup of the calculations since the costly long-range interactions can then be sampled less frequently than the

cheap short-range forces. However, the size of the outer time step in the microcanonical MTS integrators^{9–12} (e.g., RESPA) cannot be taken as being very large by simply increasing the number of inner loops, even though the distant interactions are sufficiently small. A rapid energy growth occurs when the time interval between the weak force updates exceeds half of the period related to the fastest motion.^{30–32} Such growth is well-known as resonance instabilities.^{33–35}

Within the normal mode^{13–16} and Langevin-type algorithms,^{19–22} the resonance effects are suppressed by adding friction and random forces. The mollified impulse schemes^{17–20} diminish the multistep artifacts by modifying the potential energy. In the canonical ensemble, the appearance of the resonance phenomena can be postponed to larger steps by exploiting extra phase-space variables related to a Nosé–Hoover chain thermostat.^{23–26} Alternatively, the non-Hamiltonian equations of motion which are free from the resonance instabilities can be obtained with the help of the isokinetic ensemble.²⁷ Relatively recently, it has been shown^{28,29} that a very efficient elimination of the resonant modes is achieved by conjugating the canonical chain method²³ with the isokinetic dynamics.²⁷ The resulting impulsive isokinetic Nosé–Hoover chain RESPA (INR) algorithm was applied to MD simulations of water to prove that time steps on the order of 100 fs are possible.

Despite this, the INR integrator was designed for fluids with only translational degrees of freedom. Usually, the orientational degrees are implicitly parametrized by atomic Cartesian coordinates subject to intramolecular constraints.^{36,37} However, to satisfy them within the isokinetic integrators, the cumbersome Shake- or Rattle-like iterative procedures must be involved.²⁷ The necessity to extend the MTS consideration to rotational motion is motivated by the fact that the standard force fields treat

Received: March 4, 2011

Published: November 16, 2011

water and other solvent molecules as rigid bodies.³⁸ Moreover, in a protein, it is useful to model hydrogen-containing groups as rigid moieties since the corresponding links have the highest frequencies in the molecule. The lower frequency bonds can be interpreted as flexible. Clearly, longer time steps can then be employed. Note that the rigid-body approximation yields enough accurate results without affecting the physically important distribution functions.³⁹

The existing rotational motion MD algorithms in the micro-canonical,^{40–47} canonical,^{48–50} isokinetic,^{51,52} and Langevin⁵³ ensembles are applicable solely for simple single time step (STS) dynamics, and they cannot handle much more complicated MTS integration. Surprisingly, up to now, only one paper²⁶ dealt, in fact, with the construction of MTS schemes for the propagation of orientational variables. However, it has been devoted to a canonical scheme with no emphasis on overcoming the resonance instabilities. No rigid-body MTS algorithm was derived within the isokinetic approach, which appears to be more efficient^{27–29} than the canonical method. The MTS dynamics in the presence of orientational degrees is more complex and requires a special study.

In this paper, we develop an idea of combining the isokinetic dynamics with the Nosé–Hoover-like thermostating by writing down the non-Hamiltonian equations for both translational and rotational motions. They are then explicitly integrated using the multiple time step decomposition operator method complemented by special force and torque extrapolations (section 2). This results in a completely new MTS algorithm which cannot be reduced to the impulsive INR integrator even in the absence of orientational degrees of freedom. As is demonstrated for a rigid model of water, the new approach significantly increases the maximum acceptable size of the outer step and pushes it up to several picoseconds (section 3). Concluding remarks are also provided (section 4).

2. THEORY

2.1. Interaction Site Models and Basic Equations. Let us consider a collection with N rigid molecules, each composed of M interacting sites. The usual (microcanonical) equations of translational and rotational motion for such a system can be cast in the form $d\mathbf{\Gamma}/dt = L\mathbf{\Gamma}(t)$, where⁴⁶

$$L = \sum_{i=1}^N \left[\mathbf{V}_i \frac{\partial}{\partial \mathbf{R}_i} + \mathbf{W}(\mathbf{\Omega}_i) \mathbf{S}_i : \frac{\partial}{\partial \mathbf{S}_i} + \frac{\mathbf{F}_i}{\mu} \times \frac{\partial}{\partial \mathbf{V}_i} + \mathbf{J}^{-1} (\mathbf{G}_i + (\mathbf{J}\mathbf{\Omega}_i) \times \mathbf{\Omega}_i) \frac{\partial}{\partial \mathbf{\Omega}_i} \right] \quad (1)$$

is the Liouville operator and $\mathbf{\Gamma} = \{\mathbf{R}, \mathbf{V}, \mathbf{S}, \mathbf{\Omega}\}$ denotes the set of all phase variables. Here, \mathbf{R}_i and \mathbf{V}_i are the translational position and velocity, respectively, of the center of mass $\mu = \sum_{a=1}^M m_a$ of the i th molecule, while \mathbf{S}_i and $\mathbf{\Omega}_i$ are correspondingly its attitude matrix and principal angular velocity. The matrix $\mathbf{W}(\mathbf{\Omega})$ is skewsymmetric and linear in $\mathbf{\Omega}$ so that, for instance, $\mathbf{W}(\mathbf{\Omega})\mathbf{J}\mathbf{\Omega} = (\mathbf{J}\mathbf{\Omega}) \times \mathbf{\Omega}$ with \mathbf{J} being the molecular matrix of moments of inertia.

The total force exerted on molecule i due to the site–site interactions φ_{ab} can be expressed in terms of the atomic counterparts $\mathbf{F}_{ia} = -\sum_{j \neq i}^N \sum_{b=1}^M \hat{\mathbf{r}}_{ij}^{ab} \varphi'_{ab}(r_{ij}^{ab})$ as $\mathbf{F}_i = \sum_{a=1}^M \mathbf{F}_{ia}$ where $\hat{\mathbf{r}}_{ij}^{ab} = (\mathbf{r}_{ia} - \mathbf{r}_{jb})/r_{ij}^{ab}$ with $r_{ij}^{ab} = |\mathbf{r}_{ia} - \mathbf{r}_{jb}|$ and $\varphi'(r) = d\varphi(r)/dr$. Then, the principal torque is $\mathbf{G}_i = \mathbf{S}_i \sum_{a=1}^M (\mathbf{r}_{ia} - \mathbf{R}_i) \times \mathbf{F}_{ia} = \sum_{a=1}^M \mathbf{\delta}_a \times (\mathbf{S}_i \mathbf{F}_{ia})$, where $\mathbf{r}_{ia} = \mathbf{R}_i + \mathbf{S}_i^\dagger \mathbf{\delta}_a$ denotes the position of atom a within molecule i , while $\mathbf{\delta}_a$ is the time-independent location of

site a in the body-fixed frame, so that $\mathbf{J} = \sum_{a=1}^M [(\mathbf{\delta}_a \times \mathbf{\delta}_a) \mathbf{I} - \mathbf{\delta}_a \mathbf{\delta}_a] m_a$. The functions φ_{ab} present the sum of the Lennard-Jones $4\epsilon_{ab} [(\sigma_{ab}/r_{ij}^{ab})^{12} - (\sigma_{ab}/r_{ij}^{ab})^6]$ and Coulombic $q_a q_b / r_{ij}^{ab}$ potentials, where q_a is the atomic charge. The values for parameters ϵ_{ab} , σ_{ab} , q_a and $\mathbf{\delta}_a$ depend on the concrete interaction site model chosen to describe a fluid. In the case of water, the most popular is the rigid TIP4P⁵⁷ one with $M = 4$ sites.

Because in the rigid-body approximation the intramolecular degrees of freedom are frozen, we will have only two scales of time. This follows from the fact that the total intermolecular forces $\mathbf{F} = \mathbf{F}_s + \mathbf{F}_w$ and torques $\mathbf{G} = \mathbf{G}_s + \mathbf{G}_w$ consist of the strong (s) and weak (w) parts related to the short- and long-range interactions which cause the fast and slow processes, respectively.

2.2. Standard MTS Decomposition Method. In the standard MTS decomposition approach,^{10,11} the Liouville operator is split as $L = A + B_s + B_w$ into one kinetic A and two potential $B_{s,w}$ terms. Taking into account eq 1, the explicit expressions for them are $A = \mathbf{V} \times \partial/\partial \mathbf{R} + \mathbf{W}(\mathbf{\Omega}) \mathbf{S} \partial/\partial \mathbf{S}$ as well as $B_s = \mu^{-1} \mathbf{F}_s \times \partial/\partial \mathbf{V} + \mathbf{J}^{-1} (\mathbf{G}_s + (\mathbf{J}\mathbf{\Omega}) \times \mathbf{\Omega}) \times \partial/\partial \mathbf{\Omega}$ and $B_w = \mu^{-1} \mathbf{F}_w \times \partial/\partial \mathbf{V} + \mathbf{J}^{-1} \mathbf{G}_w \times \partial/\partial \mathbf{\Omega}$, where subindex i has been omitted to simplify notation and the strong-weak components of \mathbf{F} and \mathbf{G} have been used. Then, the time evolution propagator e^{Lh} is factorized into analytically integrable single-exponential operators as $e^{[A + B_s + B_w]h} = e^{B_w(h/2)} [e^{B_s(h/2n)} e^{A(h/n)} e^{B_s(h/2n)}]^n e^{B_w(h/2)}$, where h is the size of the outer time step, n is the number of inner loops, and $\mathcal{O}(h^2)$ is the second-order error function. For any time t , the solution can be presented in the form

$$\Gamma(t) = e^{Lt} \Gamma(0) = [e^{B_w \frac{t}{2}} [e^{B_s \frac{t}{2n}} e^{A \frac{t}{n}} e^{B_s \frac{t}{2n}}]^n e^{B_w \frac{t}{2}}] \Gamma(0) + \mathcal{O}(h^2) \quad (2)$$

where $\mathcal{O}(h^2) \sim l \mathcal{L}(h^2)$ denotes the global error and $l = t/h$ is the total number of steps.

From eq 2 at $n = 1$, one reproduces the well-known Verlet integrator,^{54,55} while for $n \geq 2$, we come to the RESPA scheme.^{10,11} In the latter case, the reference system ($L_{\text{ref}} = A + B_s$) is integrated with a time step that is h/n smaller than h related to the weak contribution B_w . This speeds up the calculations because at $n > 1$ the expensive long-range forces are recalculated not so frequently. The action of the exponentials $e^{A(h/n)}$, $e^{B_s(h/(2n))}$, and $e^{B_w(h/2)}$ on a phase space point $\mathbf{\Gamma}$ can be given analytically.⁴⁶ In particular,

$$e^{A \frac{h}{n}} \{\mathbf{R}, \mathbf{S}\} = \left\{ \mathbf{R} + \frac{\mathbf{V} h}{n}, \mathbf{\Theta} \left(\mathbf{\Omega}, \frac{h}{n} \right) \mathbf{S} \right\} \quad (3)$$

where the changes of \mathbf{R} and \mathbf{S} correspond to free translational and orientational motions at fixed \mathbf{V} and $\mathbf{\Omega}$. The matrix $\mathbf{\Theta}(\mathbf{\Omega}, h/n) = \exp(\mathbf{W}(\mathbf{\Omega})h/n)$ denotes the three-dimensional rotation around vector $\mathbf{\Omega}$ on angle $\Omega h/n$.

As was mentioned in the Introduction, the RESPA scheme is hampered by the resonance instabilities already at relatively small values of h , even through the reference system is integrated exactly (i.e., when $n \gg 1$). We will now study the problem of how to eliminate these instabilities in the most efficient way.

2.3. Extrapolative Isokinetic Nosé–Hoover Chain Approach. **2.3.1. Non-Hamiltonian Equations of Motion.** Within the isokinetic ensemble,²⁷ the MTS instabilities are prevented by fixing the total kinetic energy T of the system (in our case $T = \sum_{i=1}^N [\sum_{\alpha}^{xyz} \mu V_{i\alpha}^2 / 2 + \sum_{\alpha}^{XYZ} J_{\alpha} \Omega_{i\alpha}^2 / 2]$, where J_{α} are the diagonal elements of \mathbf{J}). This energy is a collective quantity which depends on velocities of all particles. Thus, further improvements in stability

can be achieved by introducing a complete set of kinetic constraints, each one concerning only a particular degree of freedom. Obviously, such an approach should provide a better suppression of the resonant modes because it allows one to control individual kinetic energies.

The main idea consists of coupling each physical degree with its own one-dimensional, one-particle imaginary subsystem. Like real bodies, such subsystems can be characterized by some masses m and moments of inertia j_α as well as by the translational $v_{i,\alpha}$ and angular $w_{i,\alpha}$ velocities. Then, the desired individual constraints can be built in the extended phase space by writing

$$\begin{aligned} T_{i,\alpha}^t &= \mu V_{i,\alpha}^2/2 + m v_{i,\alpha}^2 = k_B \mathcal{J}/2 \\ T_{i,\alpha}^r &= J_\alpha \Omega_{i,\alpha}^2/2 + j_\alpha w_{i,\alpha}^2 = k_B \mathcal{J}/2 \end{aligned} \quad (4)$$

where k_B denotes the Boltzmann constant, \mathcal{J} is the required temperature, and α relates either to three Cartesian (x,y,z) or principal (X,Y,Z) components for the cases of translational or angular velocities. In view of eq 4, the virtual bodies can be treated as external baths or thermostats which do not allow one to exceed the fixed level $k_B \mathcal{J}/2$ of the kinetic energy for any real degree of freedom. The quantities $\mu V_{i,\alpha}^2/2$ and $J_\alpha \Omega_{i,\alpha}^2/2$ will fluctuate within the interval $[0, k_B \mathcal{J}/2]$ due to the physical interactions and energy exchange between the real system and thermostats.

The nonholonomic relations (eq 4) to be satisfied require the introduction of constraint forces $-\lambda_{i,\alpha}^t \partial T_{i,\alpha}^t / \partial V_{i,\alpha} = -\lambda_{i,\alpha}^t \mu V_{i,\alpha}$ and torques $-\lambda_{i,\alpha}^r \partial T_{i,\alpha}^r / \partial \Omega_{i,\alpha} = -\lambda_{i,\alpha}^r J_\alpha \Omega_{i,\alpha}$ for the physical system as well as their counterparts $-\lambda_{i,\alpha}^t m v_{i,\alpha}$ and $-\lambda_{i,\alpha}^r j_\alpha w_{i,\alpha}$ for the virtual degrees. Further, each such subsystem can be in turn coupled with its own chain of \mathcal{M} thermostats, described by the velocity variables $v_{j,i,\alpha}$ and $w_{j,i,\alpha}$ where $j = 1, 2, \dots, \mathcal{M}$ with $v_{i,\alpha} \equiv v_{1,i,\alpha}$ and $w_{i,\alpha} \equiv w_{1,i,\alpha}$. Acting now in the spirit of the Nosé–Hoover (NH) chain approach²³ and taking into account the constraint forces and torques, the equations of motion for the thermostat variables can be cast in the form $dv_{1,\alpha}/dt = -\lambda_{i,\alpha}^t v_{1,\alpha} - v_{1,\alpha} v_{2,\alpha}$ and $dv_{j,\alpha}/dt = (v_{j-1,\alpha}^2 - 1/\tau_t^2) - v_{j,\alpha} v_{j+1,\alpha}$ as well as $dw_{1,\alpha}/dt = -\lambda_{i,\alpha}^r w_{1,\alpha} - w_{1,\alpha} w_{2,\alpha}$ and $dw_{j,\alpha}/dt = (w_{j-1,\alpha}^2 - 1/\tau_r^2) - w_{j,\alpha} w_{j+1,\alpha}$ for $j = 2, \dots, \mathcal{M}$ with $v_{\mathcal{M}+1} = w_{\mathcal{M}+1} = 0$. Here, subscript i was again hidden for simplicity. The four relaxation times τ_t and τ_r will determine the strength of coupling of the system with the translational and rotational thermostats, so that $m = \tau_t^2 k_B \mathcal{J}/2$ and $j_\alpha = \tau_r^2 k_B \mathcal{J}/2$. This is justified by the fact that we have only one mass μ while three ($\alpha = X,Y,Z$) moments J_α of inertia of the molecule.

The Lagrangian multipliers $\lambda_{i,\alpha}^t$ and $\lambda_{i,\alpha}^r$ can be found by differentiating eq 4 with respect to time, i.e., $dT_{i,\alpha}^{t,r}/dt = 0$, and using the above equations of motion for thermostat variables complemented by the equations $dV_{i,\alpha}/dt = \mu^{-1} F_{i,\alpha} - \lambda_{i,\alpha}^t V_{i,\alpha}$ and $d\Omega_{i,\alpha}/dt = J_\alpha^{-1} (G_{i,\alpha} + (J_\beta - J_\gamma) \Omega_{i,\beta} \Omega_{i,\gamma}) - \lambda_{i,\alpha}^r \Omega_{i,\alpha}$ for translational and angular velocities. This yields

$$\begin{aligned} \lambda_{i,\alpha}^t &= \left(V_{i,\alpha} F_{i,\alpha} - \frac{1}{2} k_B \mathcal{J} \tau_t^2 v_{1,i,\alpha}^2 v_{2,i,\alpha} \right) / (2T_{i,\alpha}^t) \\ \lambda_{i,\alpha}^r &= \left(\Omega_{i,\alpha} G_{i,\alpha} + (J_\beta - J_\gamma) \Omega_{i,\alpha} \Omega_{i,\beta} \Omega_{i,\gamma} \right. \\ &\quad \left. - \frac{1}{2} k_B \mathcal{J} \tau_r^2 w_{1,i,\alpha}^2 w_{2,i,\alpha} \right) / (2T_{i,\alpha}^r) \end{aligned} \quad (5)$$

Note that the kinetic constraints (eq 4) provide true canonical distributions in position space (Appendix A).

The isokinetic Nosé–Hoover chain (INC) equations of translational and rotational motion we just derived can be

rewritten in the compact form

$$\frac{d\Gamma_{\text{inc}}}{dt} = L_{\text{inc}} \Gamma_{\text{inc}}(t) \quad (6)$$

with L_{inc} being the INC Liouville operator and $\Gamma_{\text{inc}} = \{\mathbf{R}, \mathbf{V}, \mathbf{S}, \boldsymbol{\Omega}; \mathbf{v}, \mathbf{w}\}$ denoting the extended phase space, where vectors \mathbf{v} and \mathbf{w} represent the whole set of scalar quantities $v_{j,i,\alpha}$ and $w_{j,i,\alpha}$ (subscript i will further be omitted at all). Applying the strong–weak force $\mathbf{F} = \mathbf{F}_s + \mathbf{F}_w$ and torque $\mathbf{G} = \mathbf{G}_s + \mathbf{G}_w$ decompositions, one finds in view of eq 5 that $L_{\text{inc}} = A + B_s + B_w + B_{\text{inc}}$ where now

$$\begin{aligned} B_{s,w} &= \sum_{\alpha}^{x,y,z} B_{s,w,\alpha}^t + \sum_{\alpha}^{X,Y,Z} B_{s,w,\alpha}^r \\ &= \sum_{\alpha}^{x,y,z} F_{s,w,\alpha} \left[\left(\frac{1}{\mu} - \frac{V_{\alpha}^2}{2T_{\alpha}^t} \right) \frac{\partial}{\partial V_{\alpha}} - \frac{V_{\alpha} v_{1,\alpha}}{2T_{\alpha}^t} \frac{\partial}{\partial v_{1,\alpha}} \right] \\ &\quad + \sum_{\alpha}^{X,Y,Z} (G_{s,w,\alpha} + \zeta_{s,w} (J_\beta - J_\gamma) \Omega_{\beta} \Omega_{\gamma}) \left[\left(\frac{1}{J_{\alpha}} - \frac{\Omega_{\alpha}^2}{2T_{\alpha}^r} \right) \frac{\partial}{\partial \Omega_{\alpha}} \right. \\ &\quad \left. - \frac{\Omega_{\alpha} w_{1,\alpha}}{2T_{\alpha}^r} \frac{\partial}{\partial w_{1,\alpha}} \right] \end{aligned} \quad (7)$$

at $\zeta_s = 1$ and $\zeta_w = 0$, while (α, β, γ) designate the three cyclic permutations (X,Y,Z) , (Y,Z,X) , and (Z,X,Y) . The chain thermostat contribution is

$$\begin{aligned} B_{\text{inc}} &= \sum_{\alpha}^{x,y,z} [B_{V,v,\alpha} + \sum_{j=2}^{\mathcal{M}} B_{v_j,\alpha}] \\ &\quad + \sum_{\alpha}^{X,Y,Z} [B_{\Omega,w,\alpha} + \sum_{j=2}^{\mathcal{M}} B_{w_j,\alpha}] \\ &= \sum_{\alpha}^{x,y,z} B_{\text{inc},\alpha}^t + \sum_{\alpha}^{X,Y,Z} B_{\text{inc},\alpha}^r \end{aligned} \quad (8)$$

with

$$B_{V,v,\alpha} = \frac{1}{2} V_{\alpha} v_{1,\alpha}^2 v_{2,\alpha} \tau_t^2 \frac{\partial}{\partial V_{\alpha}} + \left(\frac{1}{2} v_{1,\alpha}^3 v_{2,\alpha} \tau_t^2 - v_{1,\alpha} v_{2,\alpha} \right) \frac{\partial}{\partial v_{1,\alpha}} \quad (9)$$

and

$$B_{v_j,\alpha} = \left(v_{j-1,\alpha}^2 - \frac{1}{\tau_t^2} - v_{j,\alpha} v_{j+1,\alpha} \right) \frac{\partial}{\partial v_{j,\alpha}} \quad (10)$$

The expressions for $B_{\Omega,w,\alpha}$ and $B_{w_j,\alpha}$ are similar to those of eqs 9 and 10 with formal replacement of V by Ω , v by w , and τ_t by τ_r .

2.3.2. Extrapolative Decomposition of the Evolution Operator. Taking into account eqs 7–10, the solution to the non-Hamiltonian equations of motion (eq 6) can be obtained with the help of the decomposition method (section 2.2). As a result, for the MTS integration ($n \geq 2$), one finds

$$\Gamma_{\text{inc}}(t) = [e^{L_{\text{inc}} \frac{h}{n}} e^{B_{\text{inc}} \frac{h}{2n}} e^{B_{s,w} \frac{h}{2n}} e^{A \frac{h}{n}} e^{B_{s,w} \frac{h}{2n}} e^{B_{\text{inc}} \frac{h}{2n}}]^{n-2} e^{L_{\text{inc}} \frac{h}{n}} \Gamma_{\text{inc}}(0) + \mathcal{O}(h^2) \quad (11)$$

with

$$e^{L_{\text{inc}} \frac{h}{n}} = e^{B_{\text{inc}} \frac{h}{2n}} e^{B_{s,w} \frac{h}{2n}} e^{A \frac{h}{n}} e^{B_{s,w} \frac{h}{2n}} e^{B_{\text{inc}} \frac{h}{2n}}, \quad e^{L_{\text{inc}} \frac{h}{n}} = e^{B_{\text{inc}} \frac{h}{2n}} e^{B_{s,w} \frac{h}{2n}} e^{A \frac{h}{n}} e^{B_{s,w} \frac{h}{2n}} e^{B_{\text{inc}} \frac{h}{2n}} \quad (12)$$

where $B_{sw}^I = B_s + nB_w^I$ and $B_{sw}^{II} = B_s + nB_w^{II}$. The time evolution propagation given by eqs 11 and 12 presents an analog of the RESPA scheme (eq 2) in the case of the INC dynamics. Like RESPA, it corresponds to an impulsive approach, where the translational and angular velocities are updated instantaneously two times per outer step h (at its beginning and its end) by $e^{B_{sw}^I(h/2)}$ and $e^{B_{sw}^{II}(h/2)}$ in the weak force $F_{w,\alpha}^{I,II}$ and torque $G_{w,\alpha}^{I,II}$ fields (B_w depends on them according to eq 7). The indexes I and II mean that the forces and torques are calculated at two spatial configurations $\{\mathbf{R}(t'), \mathbf{S}(t')\}$ and $\{\mathbf{R}(t'+h), \mathbf{S}(t'+h)\}$ corresponding to two consecutive moments of time $t' = l'h$ and $t' + h$, where $l = 1, 2, \dots, l$. The arising resonance instabilities are damped out here by the thermostat stabilizing terms $e^{B_{inc}(h/(2n))}$. It is obvious, however, that with increasing the size of the time step h , the resonance effects will grow appreciably. This can prevent their proper suppressing even at a sufficiently large strength of coupling of the system with the thermostating baths.

A more efficient stabilization can be achieved within an extrapolative approach. Indeed, the smoothly varying long-range weak forces $F_{w,\alpha}$ and torques $G_{w,\alpha}$ we can hold constants $F_{w,\alpha}^I$ and $G_{w,\alpha}^I$ during the outer time interval $[t', t' + h]$. Further, such constants can be added to the quickly changing strong components $F_{s,\alpha}$ and $G_{s,\alpha}$ when performing the inner step propagation. Then, the time evolution (eqs 11 and 12) transforms to

$$\Gamma_{inc}(t) = [e^{B_{inc} \frac{h}{2n}} e^{B_{ws} \frac{h}{2n}} e^{A_n^h} e^{B_{ws} \frac{h}{2n}} e^{B_{inc} \frac{h}{2n}}]^{nl} \Gamma_{inc}(0) + \mathcal{O}(h^2) \quad (13)$$

where $B_{ws} = B_s + B_w^I$. This avoids the resonance effects since the periodic impulsive propagations of velocities by $e^{B_{sw}^{I,II}(h/2)}$ are now absent. In eq 13, the velocities are updated continuously ($2n$ times per step h) by $e^{B_{ws}^I(h/(2n))}$ in the constant weak force-torque fields $F_{w,\alpha}^I$ and $G_{w,\alpha}^I$. The next constant values $F_{w,\alpha}^{II}$ and $G_{w,\alpha}^{II}$ will be used only after passing the current outer step.

Of course, the extrapolative method produces some uncertainties to the full energy since the resulting approximate forces $F_{ws,\alpha} = F_{s,\alpha} + F_{w,\alpha}^I$ and torques $G_{ws,\alpha} = G_{s,\alpha} + G_{w,\alpha}^I$ cannot be related to any conservative potentials. However, such uncertainties have a nonresonance nature, and they are suppressed by $e^{B_{inc}(h/(2n))}$ using the same INC thermostating baths (eqs 8–10), as in the case of the impulsive method. As a result, larger values of h can be employed, despite the fact that the extrapolative algorithm (eq 13) is not reversible in time. However, this is not so important in our case because we deal with the non-Hamiltonian dynamics, where (unlike the microcanonical ensemble) the total energy should not be conserved exactly.

Note that the idea of using the force extrapolation is not new. It was exploited earlier in the context of Langevin normal mode integrators.^{13–16} However, its combination with the isokinetic Nosé–Hoover chain approach complemented with the decomposition operator method is proposed here for the first time. Moreover, the introduced constant extrapolation of the long-range torque cannot be obtained by simply fixing the atomic forces, because it depends on the orientation of the molecule as well. As will be shown in section 3, the usual atomic force extrapolation does not lead to stable phase trajectories.

Therefore, the propagation $\Gamma_{inc}(t)$ of phase variables from their initial values $\Gamma_{inc}(0)$ to arbitrary time t in the future can be readily carried out by consecutively applying the exponential operators $e^{A(h/n)}$ (see eq 3), $e^{B_{inc}(h/(2n))}$, and $e^{B_{ws}^I(h/(2n))}$ in the order defined by eq 13. The analytical expressions for the action of $e^{B_{inc}(h/(2n))}$ and $e^{B_{ws}^I(h/(2n))}$ on Γ_{inc} can be found in Appendix B.

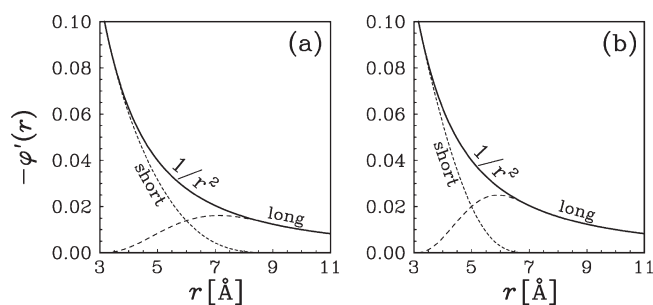


Figure 1. Schematic representation of the short- and long-range parts of the Coulombic interaction for the cutoff radii $r_c = 9 \text{ \AA}$ [subset a] and $r_c = 7 \text{ \AA}$ [subset b].

This completes the extrapolative isokinetic NH chain orientational (EINO) motion MTS algorithm. The impulsive version (eqs 11 and 12) will be referred to as simply INO. The latter was originally introduced in refs 63 and 64. In the absence of orientational degrees of freedom, the INO integrator reduces to INR.^{28,29} This is contrary to the proposed EINO approach, which cannot be reproduced from the impulsive INR scheme even in the case of pure translational motion.

3. APPLICATION TO WATER AND DISCUSSION

3.1. Details of MD Simulations. The EINO algorithm derived in the preceding section will now be tested in MD simulations. The system considered is the rigid TIP4P model⁵⁷ of water ($M = 4$). We have involved $N = 512$ molecules placed in a cubic box of volume $V = L^3$ with periodic boundaries. The simulations were performed at a density of $N/V = 1 \text{ g/cm}^3$ and a temperature of $T = 293 \text{ K}$. The total intermolecular forces \mathbf{F} were evaluated with the help of the Ewald summation technique⁵⁸ at the cutoff radii $R_c = L/2 = 12.417 \text{ \AA}$ and $\kappa_{\max} = 8$ in the real and reciprocal spaces, respectively.

The strong component \mathbf{F}_s has been determined in the spirit of the near/far distance-based approach^{11,59–62} using the replacement of $\phi'(r)$ by $\phi'_s(r) = \phi(r) \phi'(r)$ in the standard expression for \mathbf{F} (section 2.1). The switching function was chosen in the form of the cubic spline⁶² $\phi(r) = 1 - (10 - 15\eta + 6\eta^2)\eta^3$ with $\eta = 1 + (r - r_c)/(r_0 - r_c)$ to smoothly change its value from 1 to 0 when increasing the interatomic distance r from r_0 to $r_c < L/2$. The weak force part was then found by extracting \mathbf{F}_s from \mathbf{F} , i.e., as $\mathbf{F}_w = \mathbf{F} - \mathbf{F}_s$. This appears to be more efficient than the straightforward real/reciprocal splitting of Coulombic interactions.^{60,62} Having \mathbf{F}_s and \mathbf{F}_w , the strong \mathbf{G}_s and weak \mathbf{G}_w components of the total intermolecular torques \mathbf{G} were obtained applying usual relations (section 2.1) with formal replacement of \mathbf{F} by \mathbf{F}_s or \mathbf{F}_w . Two cases related to cutting-off of the short-range interaction at $r_c = 9 \text{ \AA}$ and $r_c = 7 \text{ \AA}$ have been considered. For illustration, the related short- and long-range parts $\phi'_s(r)$ and $\phi'(r) - \phi'_s(r)$ are plotted in Figure 1 together with the total function $\phi'(r)$, where $\phi(r) = 1/r$ is the generic Coulomb potential. The switching-on parameter was set equal to $r_0 = 3 \text{ \AA}$.

The equations of motion were solved using the proposed EINO algorithm (eqs 13 and B1–B8) as well as its impulsive version INO (eqs 11 and 12). For the purpose of comparison, the best previously known MTS approaches in different statistical ensembles were applied, too. They include the energy-targeted microcanonical RESPA (ERESPA) scheme as well as the extended canonical NH (ENH) and isokinetic (ISO) integrators.

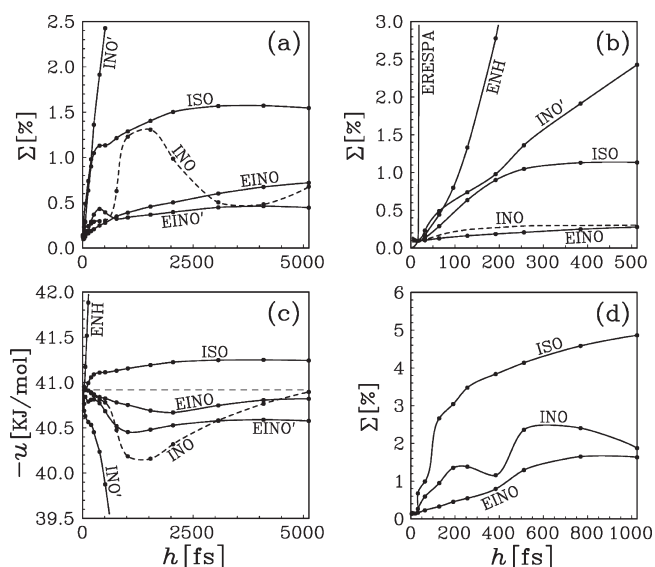


Figure 2. Uncertainty $\Sigma(h)$ in the calculation of the distribution functions by the MD simulations of the TIP4P water with different algorithms at various time steps h for $r_c = 9 \text{ \AA}$ [subsets a and b] and $r_c = 7 \text{ \AA}$ [subset d]. The mean potential energy $u(h)$ is plotted versus h in c for $r_c = 9 \text{ \AA}$.

The ERESPA, ENH, and ISO algorithms are described in detail in refs 63 and 64. Each MD run corresponded to its own size h of the outer time step. This size varied from run to run in a wide region from 4 fs up to 5120 fs. The inner step was fixed to $h/n = 4$ fs in all of the cases, meaning that the MTS parameter n changed from 1 at $h = 4$ fs to 1280 for $h = 5120$ fs. The choice $h/n = 4$ fs was dictated by the strength of the short-range intermolecular interactions. The total number l of outer steps was chosen in such a way as to cover nearly the same full propagation time $t = lh \sim 15$ ns at each given h .

During the EINO and INO propagations, we employed the three chains ($\mathcal{M} = 3$) in the extended phase space at the relaxation time $\tau_t = \tau_{r,\alpha} = \tau = 10$ fs. The triple concatenation with $n_t = n_\alpha = 8$ was applied when integrating the thermostating variables (Appendix B). The runs for $\tau = 40$ and 400 fs without concatenation at $n_t = n_\alpha = 1$ were examined too.

3.2. Numerical Results. The accuracy of the MD simulations was estimated by measuring the deviations of the oxygen–oxygen (OO), hydrogen–hydrogen (HH), and oxygen–hydrogen (OH) radial distribution functions $g(r)$ as well as the mean potential energy u of the system per molecule from their “exact” counterparts. The latter were precalculated using the Verlet integrator (i.e., RESPA at $n = 1$) with a tiny time step of $h = 1$ fs and a long simulation length of $l = t/h = 10^6$ to make the uncertainties negligibly small.^{63,64} The normalized sum (multiplied on 1/3) of the three relative root-mean-square deviations $\Sigma = (\int_0^{L/2} [g(r) - g_0(r)]^2 dr / \int_0^{L/2} g_0^2(r) dr)^{1/2}$ of $g(r)$ from the “exact” counterparts $g_0(r)$, related to the OO, HH, and OH distribution functions, is presented in Figure 2a for the ENH, ISO, INO, and EINO algorithms, depending on the size h of the outer time step. A more detailed behavior of $\Sigma(h)$ at not very large h is shown in subset b of Figure 2. The function of u on h is plotted in Figure 2c for each of the integrators (the “exact” level there is marked by the horizontal dashed line). The results in subsets a, b, and c correspond to the case $r_c = 9 \text{ \AA}$, while the dependencies of Σ on h at $r_c = 7 \text{ \AA}$ are given in Figure 2d.

We see in Figure 2 that all of the curves for $\Sigma(h)$ or $u(h)$ start at small $h \sim 4$ fs with almost the same values, which are very close to their “exact” counterparts $\Sigma(0) = 0$ or $u(0) = -40.9$ kJ/mol (a slight discrepancy is due to statistical noise). However, the further behavior of $\Sigma(h)$ and $u(h)$ strongly depends on the type of the algorithm. For example, with increasing h , the microcanonical ERESPA integrator quickly loses stability, so that already at $h \geq 20$ fs it is absolutely inadequate (see the almost vertical curve in Figure 2b). Note that the situation with the usual Verlet and RESPA algorithms is worse yet, where the maximum workable size of the time step cannot exceed 5 and 8 fs, respectively.^{63,64} A better pattern can be observed for the canonical ENH and iso-kinetic ISO integrators. However, the best results are achieved by the proposed extrapolative EINO algorithm, which exhibits exceptional accuracy and stability. Indeed, the EINO deviations are minimal in all of the quantities investigated. For example, for $r_c = 9 \text{ \AA}$ even at $h = 512$ fs, the EINO uncertainties Σ do not exceed a level of 0.3%, which is comparable with statistical noise. The impulsive INO integrator leads to an accuracy which is similar to that of the extrapolative EINO algorithm but only at not very large steps $h \lesssim 512$ fs. At longer $h \gtrsim 512$ fs, the advantage of the EINO method over the INO scheme becomes evident (cf. Figure 2a). For $r_c = 7 \text{ \AA}$, the latter is clearly inferior to the former in the whole region of h including small step sizes (cf. Figure 2d). Here, a decrease in the maximum allowable values for h is expected because the long-range interactions are stronger than in the case $r_c = 9 \text{ \AA}$. Nevertheless, even under these conditions, the extrapolative EINO algorithm can provide good precision ($\Sigma \sim 1.5\%$) in the picosecond region $h \sim 1024$ fs.

Note that the curves marked by INO and EINO in Figure 2 correspond to the value $\tau_t = \tau_{r,\alpha} = \tau = 10$ fs. It provides a sufficiently strong coupling of the system with thermostats, because then the correlation time τ is only by a factor of 2.5 larger than the inner time step $h/n = 4$ fs. Upon increasing τ to 40 fs, the performance of the impulsive method drops drastically (see the dashed curve labeled by INO'). In particular, then the maximum acceptable outer steps reduce from $h \sim 512$ to 40 fs. At the same time, the extrapolative EINO algorithm is free of this negative feature. Even at $\tau = 400$ fs, it continues to generate stable solutions in the whole region of time steps considered (see the solid curve labeled by EINO'). In addition, the EINO uncertainties only slightly increase with increasing h , giving a possibility of using the extrapolative approach up to giant values of the outer time step on the order of $h \sim 5120$ fs $\equiv 5.12$ ps, i.e., up to the picosecond region! We see in Figure 2a that the same level $\Sigma \approx 0.45\%$ of accuracy can be reached by the EINO, INO, ISO, ENH, and EOMTS integrators at the outer time steps $h = 5120, 725, 90, 60,$ and 15 fs, respectively.

The superiority of the proposed extrapolative EINO approach over its impulsive INO counterpart can be explained by the fact that the former assumes only a spatial smoothness of the long-range interactions, which should not be necessarily small. On the other hand, the impulsive method requires both smoothness and weakness of these interactions. As a consequence, the nonresonance instabilities in the extrapolative method appear to be less sensitive to the increase of the outer time step size than the resonance effects of the impulsive scheme. This is confirmed in Figure 2, where the INO uncertainties $\Sigma(h)$ and $u(h)$, unlike the EINO ones, exhibit a resonance-like behavior with the existence of maxima and minima near certain values of h . Thus, the destabilizing resonant modes still exist in the impulsive INO method, despite the usage of the thermostats.

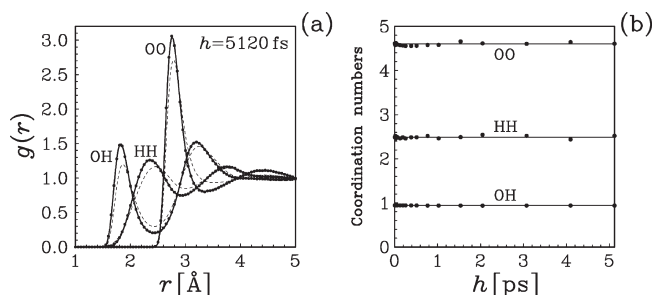


Figure 3. The OO, HH, and OH radial distribution functions of the TIP4P water obtained within the EINO algorithm for $r_c = 9 \text{ \AA}$ at $h = 5120 \text{ fs}$ [circles in a] and the corresponding coordination numbers at various time steps $h \leq 5120 \text{ fs}$ [circles in b]. The “exact” results are plotted in subsets a and b by the solid curves and horizontal lines, respectively. The dashed curves in a correspond to the microcanonical EOMTS data at $h = 19 \text{ fs}$.

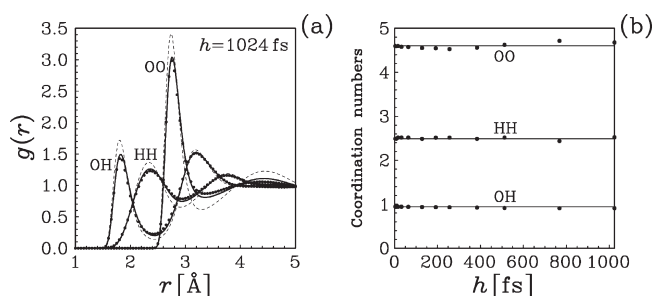


Figure 4. The same as for Figure 3 but at $r_c = 7 \text{ \AA}$ and $h \leq 1024 \text{ fs}$. The dashed curves in part a relate to a system with the long-range interactions excluded.

The OO, HH, and OH radial distribution functions $g(r)$ and their coordination numbers,^{63,64} obtained for the TIP4P water by the EINO method for $r_c = 9 \text{ \AA}$ (with $\tau = 400 \text{ fs}$) at $h = 5120 \text{ fs}$ and $h \leq 5120 \text{ fs}$, are plotted in subsets a and b of Figure 3, respectively. The “exact” results (precalculated with the help of the Verlet integrator at $h = 1 \text{ fs}$) are shown, too. The curves related to the microcanonical EOMTS scheme at $h = 19 \text{ fs}$ are included in Figure 3a as well. One can see that the deviations between the EOMTS data and the “exact” counterparts are too large already at $h = 19 \text{ fs}$. On the other hand, the EINO algorithm still produces the radial distribution functions and coordination numbers with a high level of accuracy even at a huge time step of $h = 5.12 \text{ ps}$. Indeed, the differences between the EINO and “exact” results are practically indistinguishable. Note that the largest acceptable outer time step reported earlier was $h = 100 \text{ fs}$. It has been established within the translational motion INR algorithm^{28,29} for a fully flexible water model. Thus, making the model rigid and using the proposed EINO approach have allowed us to significantly overcome this barrier. Now much larger (by a factor of 50) time step sizes on the order of 5120 fs are possible without affecting the structural properties and losing numerical stability. Such huge steps are up to 3 orders of magnitude longer than those feasible with the STS Verlet-like integrators.

The EINO distribution functions and coordination numbers corresponding to a more aggressive cutting-off with $r_c = 7 \text{ \AA}$ (and $\tau = 10 \text{ fs}$) at $h \leq 1024 \text{ fs}$ are plotted in Figure 4. The deviations here between the solid curves and circles are more visible than for

$r_c = 9 \text{ \AA}$ (cf. Figure 3) because of the increased strength of F_w and G_w . However, they are still sufficiently small to provide accurate results even at $h = 1024 \text{ fs}$. Note that the long-range interactions influence significantly even structural properties. In order to demonstrate this, the curves related to a system with no long-range forces and torques ($F_w = G_w = 0$) are also included. We see that they differ appreciably from the “exact” counterparts in the whole range of varying the interatomic distance r .

The EINO simulations in the case of the standard extrapolation, i.e., when the long-range atomic forces $F_{ia}^{(w)} = -\sum_{j \neq i} \sum_{b=1}^M (1 - \phi(r_{ij}^{ab})) \hat{r}_{ij}^{ab} \phi'_{ab}(r_{ij}^{ab})$ are held constant during outer time steps, have also been carried out. Unexpectedly, the pattern appeared to be significantly worse. It was established that already at relatively moderate $h \gtrsim 400 \text{ fs}$, the atomic force extrapolation cannot be used because then the function Σ suddenly begins to increase after $l = t/h \lesssim 400$ steps, exceeding an unacceptable level of 10%. Note that in the atomic force extrapolation the molecular torque $G_i^{(w)} = \sum_{a=1}^M \delta_a \times (S_i(t) F_{ia}^{(w)})$ is not constant and varies in time due to the reorientation $S_i(t)$ of the molecule. At first sight, such an extrapolation should look more precise since it takes into account the time dependence of $S_i(t)$. However, this will be so in the Hamiltonian dynamics at tiny values of h when the changes of $F_{ia}^{(w)}$ are small. In our case of the non-Hamiltonian propagation (which deals with huge h), it is much more important to provide a correct sampling of configurational points in phase space according to the canonical distribution (eq A1).

The molecular extrapolation just corresponds to the canonicity criteria. Indeed, the torque can be expressed in terms of the fixed (body-frame) dipole moment $\mathbf{m} = \sum_{a=1}^M \delta_a q_a$ of the molecule and the electric field $\mathbf{E}_i(t) = \sum_{j \neq i} \sum_{b=1}^M (1 - \phi(|\mathbf{R}_i - \mathbf{r}_{jb}|)) q_b (\mathbf{R}_i - \mathbf{r}_{jb}) / |\mathbf{R}_i - \mathbf{r}_{jb}|^3$ (created by all the rest of the molecules due to the long-range contribution) as $G_i^{(w)} \approx \mathbf{m} \times (S_i(t) \mathbf{E}_i(t))$. Here, we have used that $-\phi'_{ab}(r) = q_a q_b / r^2$ and $|\mathbf{r}_{ia} - \mathbf{r}_{jb}| \gg \sigma$ when calculating $F_{ia}^{(w)}$, where σ is the diameter of the molecule. Thus, the proposed extrapolation implies that the combined quantity $S_i \mathbf{E}_i$ remains fixed during the outer time step h when transiting the system from one phase space point to another. This preserves the true local canonical distribution $\exp[\mathbf{d}_i(t) \cdot \mathbf{E}_i(t) / (k_B T)] = \exp[\mathbf{m} \cdot (S_i(t) \mathbf{E}_i(t)) / (k_B T)]$ of the molecules, where $\mathbf{d}_i(t) = \sum_{a=1}^M (\mathbf{r}_{ia} - \mathbf{R}_i) q_a = S_i^\dagger(t) \mathbf{m}$ denotes the dipole moment in the laboratory frame and $-\mathbf{d}_i \cdot \mathbf{E}_i$ is the potential energy of the dipole in the electric field \mathbf{E}_i . On the other hand, the desired distribution can break significantly within the usual atomic extrapolation, where $S_i \mathbf{E}_i$ varies in time, causing the strong instabilities at large h .

Rigorously speaking, the step by step propagation of phase variables in the EINO method can be interpreted as jumps from one relevant conformation to another without going through any physical path corresponding to the original (Hamiltonian) MD. That is why such huge outer time steps of order of $h \sim 5 \text{ ps}$ are possible here. The relaxation time of reorientation of a single water molecule just belongs to the picosecond region.⁶⁵ For instance, the normalized single dipole time correlation function $\langle \mathbf{d}_i(0) \cdot \mathbf{d}_i(t) \rangle / \langle \mathbf{d}_i(0) \cdot \mathbf{d}_i(0) \rangle$ decreases from 1 at $t = 0$ to 0.25 at $t = 5 \text{ ps}$, lowering the precision of the molecular torque extrapolation when $h \gtrsim 5 \text{ ps}$ (here $\langle \dots \rangle$ denotes statistical averaging). In view of this, the step size of $h \sim 5 \text{ ps}$ should be considered as the upper theoretical limit for accurate sampling of the canonical distribution within the EINO approach.

Consider now a question on the convergence of the results. The mean potential energy u and the error Σ in the radial distributions, obtained by the EINO propagation at $\tau = 400 \text{ fs}$

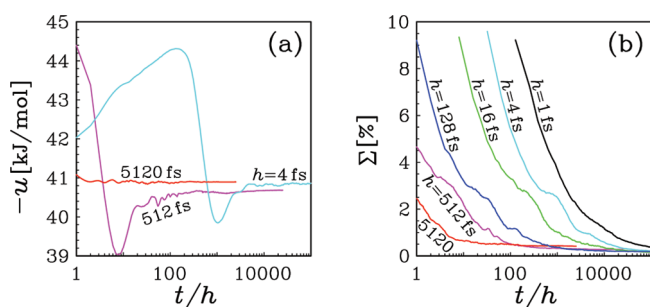


Figure 5. The EINO convergence of the potential energy [subset a] and the error in the radial distribution functions [subset b] with an increase in the length of the simulations at several fixed time steps, namely, $h = 4, 16, 128, 512$, and 5120 fs. The result in b for $h = 1$ fs corresponds to real dynamics obtained within the Verlet algorithm.

and $r_c = 9$ Å, are shown in subsets a and b of Figure 5, respectively, as functions of the number $l = t/h$ of outer steps at several characteristic values of h . In all of the runs, the initial values of phase space variables $\mathbf{\Gamma}(0)$ were taken from the same thermodynamics state pre-equilibrated at $\mathcal{T} = 293$ K but in the absence of weak long-range interactions (i.e., when $\mathbf{F}_w = \mathbf{G}_w = 0$). Then, the weak forces and torques were turned on at time $t = 0$, and the relaxation of the system from the perturbed state to the true configuration was observed. The potential energy was measured every outer time step when performing statistical averaging. Then, however, the measurements were carried out too frequently at small h , leading to a computational overload. In order to avoid this, the statistics for the radial distribution functions were accumulated after each fixed time interval of 128 fs for any h . Then, the costs due to the measurements will be negligibly small with respect to those spent on the integration of phase space variables.

We see in Figure 5a for the potential energy that the convergence at the largest stepsize $h = 5120$ fs is considerably faster than at moderate ($h = 512$ fs) and small ($h = 4$ fs) time steps. For instance, already after four outer time steps of size $h = 5120$ fs each, the potential energy almost achieves its limiting value and virtually does not change with a further increase in the simulation length. On the other hand, the asymptotic regime at $h = 512$ and 4 fs is reached only when $t/h \sim 40$ and 4×10^3 , respectively. Thus, nearly the same time interval on the order of 20 ps is necessary to obtain reliable results in all three cases. Similar behavior is inherent in the error function Σ (see Figure 5b). Here, for instance, a level of $\Sigma = 1\%$ is reached after $t/h = 1.5 \times 10^4$, 4×10^3 , 10^3 , 150, 40, and 4 outer steps at $h = 1, 4, 16, 128, 512$, and 5120 fs, respectively. Such numbers are exactly inversely proportional to h , indicating that almost the same propagation time on the order of $t \sim 20$ ps is again enough to reproduce the radial distribution functions within 1% precision for all six sizes of h . This is confirmed in Figure 6, where the EINO curves are practically indistinguishable. Moreover, they are close to those corresponding to real dynamics at $h = 1$ fs (slight deviations in Figure 6a at intermediate t are explained by relaxation of the chain variables, which were set equal to zero, $w_j = 0$ and $v_j = 0$ with $j = 2, \dots, \mathcal{M}$, at $t = 0$). Hence, the quasidynamics produced by the EINO method are free from the drag that would be caused by the introduction of the thermostats.

3.3. CPU Speedup. From the information already given, we can conclude that the EINO approach allows huge sizes of the outer time step which at $h \sim 5120$ fs are by a factor of $5120/8 = 640$ larger

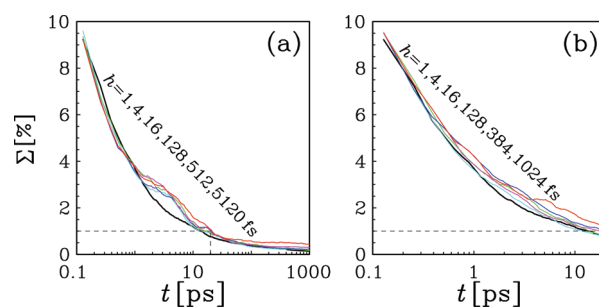


Figure 6. The EINO convergence of the error Σ in the radial distributions as a function of time t at several fixed step sizes h for $r_c = 9$ Å [$\tau = 400$ fs, subset a] and $r_c = 7$ Å [$\tau = 10$ fs, subset b]. Other notations are similar to those of Figure 5b.

than those of the standard RESPA scheme (for the latter, the maximal h is^{63,64} on the order of 8 fs). Note, however, that the ideal 640-fold increase in computational efficiency cannot be realized because a nonzero portion $\theta > 0$ of CPU time is needed to evaluate the short-range forces and torques. This portion ($0 < \theta < 1$) can be measured in terms of the cost ratio of calculating such (cheap) forces and torques to (expensive) long-range interactions. Let h_{MTS} and h_{STS} be the maximal sizes of the time step which are possible to use by some MTS ($n > 1$) and STS ($n_{\text{STS}} = 1$) algorithms, respectively. Then, taking into account that nearly the same integration time t (see section 3.2) is required to reach an asymptotic behavior for all of the approaches considered, the actual relative speedup can be estimated as $\Lambda = n(1 - \xi)(1 + \theta)/(1 + n\theta)$. Here, $n = h_{\text{MTS}}/h_{\text{STS}}$ with $h_{\text{STS}} = 4$ fs, since the fixed inner step ($h/n = 4$ fs) is employed for any n to achieve the same precision corresponding to fast dynamics. The multiplier $0 < \xi < 1$ takes into account the overhead of the INO and EINO techniques on the propagation of extra (thermostatting) phase variables.

Thus, the ideal CPU speedup $\Lambda_{\text{max}} = n$ can be expected only in a hypothetical case when $\theta \rightarrow 0$ and $\xi \rightarrow 0$. In practical calculations $\Lambda < n$ because the quantities θ and ξ are always finite. These quantities depend on details of the simulations, implemented program code, and the compilers and platform used. The present calculations were performed on the SGI Altix 4700 supercomputer using the Linux Intel Fortran compiler. In the case of $N = 512$ with $r_c = 9$ Å and $R_c = L/2 = 12.417$ Å, we found that $\theta \approx 1/20$, while $\xi \approx 0.1$. With decreasing r_c to 7 Å, the ratio θ decreases to 1:50. The CPU speedups Λ for different MTS algorithms, obtained in the MD simulations of water with respect to the STS Verlet integrator (at $h = 4$ fs), are plotted in Figure 7 as functions of the size of the outer time step by the lower ($r_c = 9$ Å) and medium ($r_c = 7$ Å) lying curves. The symbols are related to the maximal values of h , which are still safely workable within a given algorithm.

As can be seen in Figure 7, the RESPA and ERESPA schemes only slightly ($\Lambda = 2-3$) increase the efficiency. A better pattern is observed for the ENH and ISO integrators which can reduce the overall CPU costs up to $\Lambda = 10$ times. The best results are achieved within the INO (triangles) and EINO (full circles) algorithms. They are able to speed up the MD calculations by factors of $\Lambda = 20$ and 40, respectively. Such an increase is explained by larger values of h , and thus n , which are inherent in these algorithms. Remember that to cover the same integration time t , the total number of steps t/h decreases with increasing h .

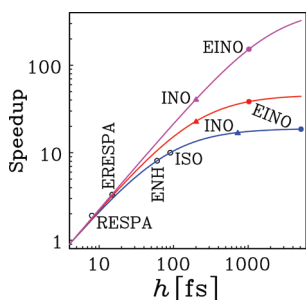


Figure 7. CPU speedup for different MTS algorithms in MD simulations of water relative to the STS integrator. The results obtained for $N = 512$ molecules with cutoff radii $r_c = 9 \text{ \AA}$ and $r_c = 7 \text{ \AA}$ appear as the lower and middle lying curves, respectively. The theoretical estimations for $N = 5120$ with $r_c = 7 \text{ \AA}$ are presented by the upper curve. The symbols correspond to the maximum allowable values of h for each of the algorithms.

This leads to an increase in the overall computational efficiency since the extra INO and EINO thermostating costs are minimal ($\xi \approx 0.1$). Note also that the values of Λ increase with decreasing the cutoff radius r_c from 9 \AA to 7 \AA because of lowering θ , despite the decrease of the maximum allowable size of the outer time step. This is so because then the computational costs spent on the calculation of the short-range interactions drop significantly. Moreover, taking into account that the ratio θ is approximately proportional to $(r_c/R_c)^3$, the speedup will further increase with an increase in the size of the system. For instance, for a collection of $N = 5120$ water molecules, it is expected that the relative speedup will be on the order of $\Lambda \approx 150$ (the upper lying curve in Figure 7). The choice of optimal values for r_c and R_c goes beyond the scope of the current paper and will be reported elsewhere.

4. CONCLUSION

We have proposed a new multiscale approach to overcome the restrictions on time step sizes in MD simulations of interaction site models of fluids with orientational degrees of freedom. It presents a nontrivial combination of the decomposition operator method with a special extrapolation of intermolecular interactions complemented by a modified isokinetic Nosé–Hoover chain thermostat. This has allowed us to substantially enlarge the size of the outer time step when propagating the phase space variables and, thus, significantly improve the efficiency of MD computations. As is shown on the basis of MD simulations for the rigid TIP4P model of ambient water, giant step sizes up to several picoseconds become possible without losing numerical stability and affecting equilibrium properties. Such steps are up to 3 orders of magnitude larger than those of single-scale Verlet-type integrators and by a factor of 50 longer than the maximal time steps feasible with the best previous multiple time step algorithms. This constitutes a considerable advantage for MD simulation of molecular liquids, with many applications in solution chemistry.

The new approach could be extended to more complex models of liquids and solutions in the presence of both rigid and flexible atomic groups, including solvated proteins and other biomolecules. The latter presents the biggest challenges, since it requires introducing three or more time scales. The fastest one is related to the internal bond and bending vibrations of atoms within the molecule. For large proteins, we should take into account very slow collective dynamics of molecular domains which can influence the atomic motion. Then, an interplay between the

solvent hydrodynamics and solute movements will also take place, resulting in an extremely large separation between the time scales. Similar difficulties might arise when coupling the proposed MD methodology with the statistical–mechanical, 3D-RISM-KH molecular theory of solvation.^{67–69} All of these problems will be a subject of future investigations.

APPENDIX A: CANONICITY OF THE INC DYNAMICS

It should be pointed out that the quasdynamics generated by the proposed INC equations of motion cannot provide the canonical distribution in velocity space because of the imposed individual kinetic constraints (eq 4). Nevertheless, the configurational part $\mathcal{Z}(\mathcal{F})$ of the extended partition function $\mathcal{G}(\mathcal{F})$ obtained within the INC approach does correspond to the true canonical distribution of the physical system (at temperature \mathcal{T}). This allows one to perform the genuine canonical averages of position- and orientation-dependent properties in equilibrium.

Indeed, taking into account eq 4, one sees that the partition function in the extended phase space (the basic system plus chain thermostats) is of the form

$$\begin{aligned} \mathcal{Q} &= \int \prod_{i=1}^N d\mathbf{v}_i d\Omega_i dv_i dw_i \\ &\times \prod_{\alpha}^{x,y,z} \delta(\mu V_{i,\alpha}^2 + m_i v_{1,i,\alpha}^2 - k_B \mathcal{T}) \\ &\times \prod_{\alpha}^{x,y,z} \delta(J_{\alpha} \Omega_{i,\alpha}^2 + j_{i,\alpha} w_{1,i,\alpha}^2 - k_B \mathcal{T}) \\ &\times \prod_{a=1}^M d\mathbf{r}_{ia} e^{(-U(\mathbf{r}) + \sum_{i,\alpha} T_{i,\alpha}(V, \Omega, v, w)) / (k_B \mathcal{T})} \\ &\propto \prod_{i,a=1}^{N,M} d\mathbf{r}_{ia} e^{-U(\mathbf{r}) / (k_B \mathcal{T})} \propto \mathcal{Z}(\mathcal{F}) \end{aligned} \quad (\text{A1})$$

where $\mathbf{r} \equiv \{\mathbf{r}_{ia}\}$ denotes the whole set of atomic positions, $U(\mathbf{r})$ is the full potential energy of the system, and $T_{i,\alpha}(V, \Omega, v, w) = \mu V_{i,\alpha}^2 / 2 + J_{\alpha} \Omega_{i,\alpha}^2 / 2 + \sum_{j=1}^M (m v_{j,i,\alpha}^2 + j_{\alpha} w_{j,i,\alpha}^2) / 2$ are the i, α th components of the total kinetic energy, which includes the real and imaginary velocity-type variables. Integrating in eq A1 over all of these variables and taking into account the presence of the δ functions gives the desired result $\mathcal{G} \propto \mathcal{Z}$.

APPENDIX B: ANALYTICAL EXPRESSIONS FOR SINGLE EXPONENTIAL OPERATORS

Apart from the possibility of using huge time steps, another great advantage of the proposed EINO method is that all of the single exponentials in eq 13 can be handled analytically. Really, in view of eq 7, the actions of operators $e^{B_{ws} h / (2n)}$ on translational phase variables \mathbf{V} and \mathbf{v}_1 can first be factorized into the Cartesian components

$$e^{B_{ws} \frac{h}{2n} \{\mathbf{V}, \mathbf{v}_1\}} = \prod_{\alpha}^{x,y,z} e^{B_{ws, \alpha}^{\frac{h}{2n}} \{V_{\alpha}, v_{1,\alpha}\}} \quad (\text{B1})$$

and then expressed in terms of hyperbolic functions

$$\begin{aligned} &e^{B_{ws, \alpha}^{\frac{h}{2n}} \{V_{\alpha}, v_{1,\alpha}\}} \\ &= \left\{ \frac{V_{\alpha} + \vartheta_{\alpha}^{-1} \tanh(h_{t,\alpha})}{1 + V_{\alpha} \vartheta_{\alpha} \tanh(h_{t,\alpha})}, \frac{v_{1,\alpha} \cosh^{-1}(h_{t,\alpha})}{1 + V_{\alpha} \vartheta_{\alpha} \tanh(h_{t,\alpha})} \right\} \end{aligned} \quad (\text{B2})$$

with $\vartheta_\alpha = (T_\alpha^t/\mu)^{-1/2}$ and $h_{t,\alpha} = \mu^{-1}F_{ws,\alpha}\vartheta_\alpha h/(2n)$. The expressions in the case of rotation motion for orientational phase variables $\mathbf{\Omega}$ and \mathbf{w}_1 are somewhat more complicated, namely,

$$e^{B_{ws,\alpha}^t \frac{h}{2n}} \{\mathbf{\Omega}, \mathbf{w}_1\} = e^{B_{ws,X}^t \frac{h}{4n}} \{\mathbf{\Omega}_X, w_{1,X}\} e^{B_{ws,Y}^t \frac{h}{4n}} \{\mathbf{\Omega}_Y, w_{1,Y}\} \\ \times e^{B_{ws,Z}^t \frac{h}{4n}} \{\mathbf{\Omega}_Z, w_{1,Z}\} e^{B_{ws,X}^t \frac{h}{4n}} \{\mathbf{\Omega}_Y, w_{1,Y}\} \\ \times e^{B_{ws,X}^t \frac{h}{4n}} \{\mathbf{\Omega}_X, w_{1,X}\} + \mathcal{O}(h^3) \quad (\text{B3})$$

where

$$e^{B_{ws,\alpha}^t \frac{h}{2n}} \{\mathbf{\Omega}_\alpha, w_{1,\alpha}\} \\ = \left\{ \frac{\Omega_\alpha + \chi_\alpha^{-1} \tanh(h_{r,\alpha})}{1 + \Omega_\alpha \chi_\alpha \tanh(h_{r,\alpha})}, \frac{w_{1,\alpha} \cosh^{-1}(h_{r,\alpha})}{1 + \Omega_\alpha \chi_\alpha \tanh(h_{r,\alpha})} \right\} \quad (\text{B4})$$

with $\chi_\alpha = (T_\alpha^t/J_\alpha)^{-1/2}$, $h_{r,\alpha} = J_\alpha^{-1}G_\alpha \chi_\alpha h/(2n)$, and $G_\alpha = G_{ws,\alpha} + (J_\beta - J_\gamma)\Omega_\beta \Omega_\gamma$. It should be mentioned that the principal ($\alpha = X, Y, Z$) components of the orientational part of $B_{ws,\alpha}$ do not commute because of the existence of the inertial torque terms $(J_\beta - J_\gamma)\Omega_\beta \Omega_\gamma$ (see eq 7). Thus, contrary to the simple factorization (eq B1) used for the Cartesian ($\alpha = x, y, z$) components of the translational part of $B_{ws,\alpha}$ the extra decomposition (eq B3) has been exploited to achieve the desired $\mathcal{O}(h^3)$ one-step accuracy. Then, the partial operators $e^{B_{ws,\alpha}^t h/(2n)}$ acting on angular velocities $\mathbf{\Omega}$ and \mathbf{w}_1 will change only their α th component according to eq B4 at constant values of the remaining two parts β and γ .

Additional splitting is needed for analytical handling of the INC thermostat propagator $e^{B_{inc}^t h/(2n)}$. In view of eqs 8–10, it can be decomposed as

$$e^{B_{inc}^t \frac{h}{2n}} = \prod_{\alpha}^{x,y,z} [e^{B_{inc,\alpha}^t \frac{h}{2nm_\alpha}}]^{n_t} \prod_{\alpha}^{X,Y,Z} [e^{B_{inc,\alpha}^t \frac{h}{2nm_\alpha^r}}]^{n_\alpha^r} \quad (\text{B5})$$

where (for $\mathcal{M} \leq 3$):

$$e^{B_{inc,\alpha}^t \frac{h}{2nm_\alpha}} = e^{B_{v_3,\alpha}^t \frac{h}{8nm_\alpha}} e^{B_{v_2,\alpha}^t \frac{h}{4nm_\alpha}} e^{B_{v_3,\alpha}^t \frac{h}{8nm_\alpha}} e^{B_{v_1,\alpha}^t \frac{h}{2nm_\alpha}} e^{B_{v_3,\alpha}^t \frac{h}{8nm_\alpha}} \\ \times e^{B_{v_2,\alpha}^t \frac{h}{4nm_\alpha}} e^{B_{v_3,\alpha}^t \frac{h}{8nm_\alpha}}, \\ e^{B_{inc,\alpha}^t \frac{h}{2nm_\alpha^r}} = e^{B_{w_3,\alpha}^t \frac{h}{8nm_\alpha^r}} e^{B_{w_2,\alpha}^t \frac{h}{4nm_\alpha^r}} e^{B_{w_3,\alpha}^t \frac{h}{8nm_\alpha^r}} e^{B_{\Omega,w,\alpha}^t \frac{h}{2nm_\alpha^r}} e^{B_{w_3,\alpha}^t \frac{h}{8nm_\alpha^r}} \\ \times e^{B_{w_2,\alpha}^t \frac{h}{4nm_\alpha^r}} e^{B_{w_3,\alpha}^t \frac{h}{8nm_\alpha^r}} \quad (\text{B6})$$

In eq B6, the internal loops with $n_t \geq 1$ and $n_\alpha^r \geq 1$ cycles have been introduced to improve the accuracy of the INC thermostat integration. This is necessary if the thermostating correlation times are small, i.e., if $\tau_t \sim h/n$ and $\tau_{r,\alpha} \sim h/n$. Then, n_t and n_α^r should be chosen in order to satisfy the inequalities $h/(2nm_\alpha) \ll \tau_t$ and $h/(2nm_\alpha^r) \ll \tau_{r,\alpha}$. The extra precision can be reached by applying the triple concatenation⁵⁶ of eq B6 at $h \equiv \zeta h$, $(1 - 2\zeta)h$, and again at $h \equiv \zeta h$, where $\zeta = 1/(2 - 2^{1/2})$. Such a concatenation reduces the uncertainty of the decompositions from $\mathcal{O}(h^3)$ to a negligibly small level of $\mathcal{O}(h^5)$. When $\tau_t \gg h/n$ and $\tau_{r,\alpha} \gg h/n$, one can set $n_t = 1$ and $n_\alpha^r = 1$.

The action of the single exponential operators in eq B6 on the extended phase variables can be presented in terms of elementary functions as well. Using eqs 9 and 10, one finds for the

translational components

$$e^{B_{v,\alpha}^t \frac{h}{2nm_\alpha}} \{V_\alpha, v_{1,\alpha}\} = \left[1 + \frac{v_{1,\alpha}^2 \tau_t^2 k_B \mathcal{F}}{4T_\alpha^t} \right. \\ \left. \times (e^{-v_{2,\alpha} \frac{h}{4nm_\alpha}} - 1) \right]^{-1/2} \{V_\alpha, e^{-v_{2,\alpha} \frac{h}{4nm_\alpha}} v_{1,\alpha}\}, \\ e^{B_{v_2,\alpha}^t \frac{h}{4nm_\alpha}} v_{2,\alpha} = v_{2,\alpha} e^{-v_{3,\alpha} \frac{h}{8nm_\alpha}} + 2 \left(v_{1,\alpha}^2 - \frac{1}{\tau_t^2} \right) \\ \times \frac{e^{-v_{3,\alpha} \frac{h}{8nm_\alpha}} \sinh \left(v_{3,\alpha} \frac{h}{8nm_\alpha} \right)}{v_{3,\alpha}}, \\ e^{B_{v_3,\alpha}^t \frac{h}{8nm_\alpha}} v_{3,\alpha} = v_{3,\alpha} + \left(v_{2,\alpha}^2 - \frac{1}{\tau_t^2} \right) \frac{h}{8nm_\alpha} \quad (\text{B7})$$

The analogous expressions for the rotational components read

$$e^{B_{\Omega,w,\alpha}^t \frac{h}{2nm_\alpha}} \{\mathbf{\Omega}_\alpha, w_{1,\alpha}\} = \left[1 + \frac{w_{1,\alpha}^2 \tau_r^2 k_B \mathcal{F}}{4T_\alpha^t} \right. \\ \left. \times (e^{-w_{2,\alpha} \frac{h}{4nm_\alpha^r}} - 1) \right]^{-1/2} \{\mathbf{\Omega}_\alpha, e^{-w_{2,\alpha} \frac{h}{4nm_\alpha^r}} w_{1,\alpha}\}, \\ e^{B_{w_2,\alpha}^t \frac{h}{4nm_\alpha^r}} w_{2,\alpha} = w_{2,\alpha} e^{-w_{3,\alpha} \frac{h}{4nm_\alpha^r}} + 2 \left(w_{1,\alpha}^2 - \frac{1}{\tau_r^2} \right) \\ \times \frac{e^{-w_{3,\alpha} \frac{h}{4nm_\alpha^r}} \sinh \left(w_{3,\alpha} \frac{h}{8nm_\alpha^r} \right)}{w_{3,\alpha}}, \\ e^{B_{w_3,\alpha}^t \frac{h}{8nm_\alpha^r}} w_{3,\alpha} = w_{3,\alpha} + \left(w_{2,\alpha}^2 - \frac{1}{\tau_r^2} \right) \frac{h}{8nm_\alpha^r} \quad (\text{B8})$$

Note that the simultaneous transformations of $(V_\alpha, v_{1,\alpha})$ and $(\mathbf{\Omega}_\alpha, w_{1,\alpha})$ given by eqs B2 and B4 as well as B7 and B8 conserve the individual kinetic constraints (eq 4) to within a machine accuracy at any time step size h . This has been achieved by the special decompositions (eqs 7 and 8) and analytical (i.e., exact) expressions for the single exponential propagators. Such a conservation must be considered as a very important feature of the proposed algorithm because now in principle arbitrarily large values of h can be exploited without a loss of numerical stability despite the fact that the phase trajectories are produced approximately ($\mathcal{O}(h^2) \neq 0$ in eq 13). The stability can be improved additionally by recalculating $T_\alpha^t = \mu V_\alpha^2/2 + k_B \mathcal{F} \tau_t^2 v_{1,\alpha}^2/4$ and $\mathcal{J} T_\alpha^t \equiv J_\alpha \Omega_\alpha^2/2 + k_B \mathcal{F} \tau_r^2 w_{r,\alpha}^2/4$ in eqs B2, B4, B7, and B8 before each time step rather to merely put $T_\alpha^t = k_B \mathcal{F}/2$ and $T_\alpha^r = k_B \mathcal{F}/2$. This prevents the accumulation of machine errors and provides the constraint conservation not only locally but also globally for any time $t \gg h$.

It should be pointed out also that the relatively large number of single exponentials appearing in the EINO propagation presents no numerical difficulties. The action of these exponentials on a phase space point leads to simple analytical transformations given by elementary functions. They incur practically negligible computational costs, compared to those necessary to spend on the calculation of intermolecular forces and torques. The numerical overheating can be reduced to a minimum by replacing the elementary (exponents and hyperbolic trigonometric) functions with their rational counterparts. This can be useful especially for thermostating propagation (eqs B5 and B6) if $n_t > 1$ and $n_\alpha^r > 1$. The rational counterparts can be obtained by expanding the

functions in power series with respect to their arguments, taking into account that the latter are small. The expansion can be restricted to a finite number of terms within the required precision. For instance, for the first lines of eq B7 and B8, when $\xi = v_{2,\alpha}h/(2m\tau)$ or $\xi = w_{2,\alpha}h/(2m\tau_{\alpha})$, we have $e^{-\xi} = (1 - \xi/2)/(1 + \xi/2) + \mathcal{O}(h^3)$ or even $e^{-\xi} = (1 - \xi/2 + \xi^2/12)/(1 + \xi/2 + \xi^2/12) + \mathcal{O}(h^5)$ that already exceeds the one-step precision $\mathcal{O}(h^3)$ of the basic integration. At the same time, in these lines, it is necessary to put $e^{-2\xi} = (e^{-\xi})^2$ rather than directly expand $e^{-2\xi}$. This maintains the exact conservation of the kinetic constraints. Similar counterparts can be used in the second lines of eqs B7 and B8 exploiting the equality $2e^{-\xi/2} \sinh(\xi/2)/\xi = (1 - e^{-\xi})/\xi = 1/(1 + (\xi/2) + \mathcal{O}(h^3))$ or $1/(1 + \xi/2 + \xi^2/12) + \mathcal{O}(h^5)$.

AUTHOR INFORMATION

Corresponding Author

*E-mail: omelyan@icmp.lviv.ua; andriy.kovalenko@nrc-cnrc.gc.ca.

ACKNOWLEDGMENT

We gratefully acknowledge the financial support by the ArboNano—the Canadian Forest NanoProducts Network and by the National Research Council (NRC) of Canada. I.P.O. is thankful for the hospitality during his stay at the University of Alberta and the National Institute for Nanotechnology.

REFERENCES

- Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Clarendon: Oxford, U.K., 1987.
- Frenkel, D.; Smit, B. *Understanding Molecular Simulation: from Algorithms to Applications*; Academic Press: New York, 1996.
- Theodorou, D. N.; Kotelianski, M. *Simulation Methods for Polymers*; Marcel Dekker: New York, 2004.
- Leimkuhler, B.; Reich, S. *Simulating Hamiltonian Dynamics*; Cambridge University Press: Cambridge, U.K., 2005.
- Rojnuckarin, A.; Kim, S.; Subramaniam, S. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 4288.
- Hernández, G.; Jenney, F. E., Jr.; Adams, M. W. W.; LeMaster, D. M. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 3166.
- Karplus, M.; McCammon, J. A. *Nat. Struct. Biol.* **2002**, *9*, 646.
- Zhang, Y.; Peters, M. H.; Li, Y. *Proteins: Struct., Funct., Genet.* **2003**, *52*, 339.
- Grubmüller, H.; Heller, H.; Windemuth, A.; Schulten, K. *Mol. Simul.* **1991**, *6*, 121.
- Tuckerman, M. E.; Berne, B. J.; Martyna, G. J. *J. Chem. Phys.* **1992**, *97*, 1990.
- Stuart, S. J.; Zhou, R.; Berne, B. J. *J. Chem. Phys.* **1996**, *105*, 1426.
- Kopf, A.; Paul, W.; Dünweg, B. *Comput. Phys. Commun.* **1997**, *101*, 1.
- Zhang, G.; Schlick, T. *J. Comput. Chem.* **1993**, *14*, 1212.
- Zhang, G.; Schlick, T. *J. Chem. Phys.* **1994**, *101*, 4995.
- Schlick, T.; Barth, E.; Mandziuk, M. *Annu. Rev. Biophys. Biomol. Struct.* **1997**, *26*, 181.
- Barth, E.; Schlick, T. *J. Chem. Phys.* **1998**, *109*, 1617.
- Garcia-Archilla, B.; Sanz-Serna, J. M.; Skeel, R. D. *SIAM J. Sci. Comput.* **1998**, *20*, 930.
- Izaguirre, J. A.; Reich, S.; Skeel, R. D. *J. Chem. Phys.* **1999**, *110*, 9853.
- Ma, Q.; Izaguirre, J. A. *Multiscale Model. Simul.* **2003**, *2*, 1.
- Izaguirre, J. A.; Catarello, D. P.; Wozniak, J. M.; Skeel, R. D. *J. Chem. Phys.* **2001**, *114*, 2090.
- Skeel, R. D.; Izaguirre, J. A. *Mol. Phys.* **2002**, *100*, 3885.
- Melchionna, S. *J. Chem. Phys.* **2007**, *127*, 044108.
- Martyna, G. J.; Tuckerman, M. E.; Tobias, D. J.; Klein, M. L. *Mol. Phys.* **1996**, *87*, 1117.
- Cheng, A.; Merz, K. M., Jr. *J. Phys. Chem. B* **1999**, *103*, 5396.
- Komeiji, J. *THEOCHEM* **2000**, *530*, 237.
- Shinoda, W.; Mikami, M. *J. Comput. Chem.* **2003**, *24*, 920.
- Minary, P.; Martyna, G. J.; Tuckerman, M. E. *J. Chem. Phys.* **2003**, *118*, 2510.
- Minary, P.; Tuckerman, M. E.; Martyna, G. J. *Phys. Rev. Lett.* **2004**, *93*, 150201.
- Abrams, J. B.; Tuckerman, M. E.; Martyna, G. J. *Computer Simulations in Condensed Matter Systems: From Materials to Chemical Biology*; Springer-Verlag: Berlin, 2006; Vol. 1. [*Lect. Notes Phys.* **2006**, *703*, 139.]
- Zhou, R.; Berne, B. J. *J. Chem. Phys.* **1995**, *103*, 9444.
- Watanabe, M.; Karplus, M. *J. Phys. Chem.* **1995**, *99*, 5680.
- Barth, E.; Schlick, T. *J. Chem. Phys.* **1998**, *109*, 1633.
- Mandziuk, M.; Schlick, T. *Chem. Phys. Lett.* **1995**, *237*, 525.
- Schlick, T.; Mandziuk, M.; Skeel, R. D.; Srinivas, K. *J. Comput. Phys.* **1998**, *140*, 1.
- Ma, Q.; Izaguirre, J. A.; Skeel, R. D. *SIAM J. Sci. Comput.* **2003**, *24*, 1951.
- Ciccotti, G.; Ryckaert, J. P.; Ferrario, M. *Mol. Phys.* **1982**, *47*, 1253.
- Andersen, H. C. *J. Comput. Phys.* **1983**, *52*, 24.
- MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L., Jr.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E., III; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586.
- Chen, B.; Martin, M. G.; Siepmann, J. I. *J. Phys. Chem. B* **1998**, *102*, 2578.
- Dullweber, A.; Leimkuhler, B.; McLachlan, R. J. *Chem. Phys.* **1997**, *107*, 5840.
- Omelyan, I. P. *Comput. Phys.* **1998**, *12*, 97.
- Omelyan, I. P. *Comput. Phys. Commun.* **1998**, *109*, 171.
- Omelyan, I. P. *Phys. Rev. E* **1998**, *58*, 1169.
- Matubayasi, N.; Nakahara, M. *J. Chem. Phys.* **1999**, *110*, 3291.
- Miller, T. F., III; Eleftheriou, M.; Pattanaik, P.; Ndirango, A.; News, D.; Martyna, G. J. *J. Chem. Phys.* **2002**, *116*, 8649.
- Omelyan, I. P. *J. Chem. Phys.* **2007**, *127*, 044102.
- Omelyan, I. P. *Phys. Rev. E* **2008**, *78*, 026702.
- Ikeguchi, M. *J. Comput. Chem.* **2004**, *25*, 529.
- Kamberaj, H.; Low, R. J.; Neal, M. P. *J. Chem. Phys.* **2005**, *122*, 224114.
- Okumura, H.; Itoh, S. G.; Okamoto, Y. *J. Chem. Phys.* **2007**, *126*, 084103.
- Kuttel, R.; Jones, R. B. *Phys. Rev. E* **2000**, *61*, 3186.
- Terada, T.; Kidera, A. *J. Chem. Phys.* **2002**, *116*, 33.
- Davidchack, R. L.; Handel, R.; Tretyakov, M. V. *J. Chem. Phys.* **2009**, *130*, 234101.
- Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. *J. Chem. Phys.* **1982**, *76*, 637.
- Omelyan, I. P.; Mryglod, I. M.; Folk, R. *Phys. Rev. E* **2002**, *65*, 056706.
- Creutz, M.; Gocksch, A. *Phys. Rev. Lett.* **1989**, *63*, 9.
- Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.
- Omelyan, I. P. *Comput. Phys. Commun.* **1997**, *107*, 113.
- Zhou, R.; Harder, E.; Xu, H.; Berne, B. J. *J. Chem. Phys.* **2001**, *115*, 2348.
- Qian, X.; Schlick, T. *J. Chem. Phys.* **2002**, *116*, 5971.
- Han, G.; Deng, Y.; Glimm, J.; Martyna, G. *Comput. Phys. Commun.* **2007**, *176*, 271.
- Morrone, J. A.; Zhou, R.; Berne, B. J. *J. Chem. Theory Comput.* **2010**, *6*, 1798.

- (63) Omelyan, I. P.; Kovalenko, A. *J. Chem. Phys.* **2011**, *135*, 114110.
- (64) Omelyan, I. P.; Kovalenko, A. *J. Chem. Phys.* **2011**, to be published.
- (65) Omelyan, I. P. *Mol. Phys.* **1998**, *93*, 123.
- (66) Omelyan, I. P.; Mryglod, I. M.; Tokarchuk, M. V. *Condens. Matter Phys.* **2005**, *8*, 25.
- (67) Kovalenko, A. In *Molecular Theory of Solvation*; Hirata, F., Ed.; Kluwer Academic Publishers: Norwell, MA, 2003; Vol. 24, Chapter 4.
- (68) Miyata, T.; Hirata, F. *J. Comput. Chem.* **2008**, *29*, 871.
- (69) Luchko, T.; Gusarov, S.; Roe, D. R.; Simmerling, C.; Case, D. A.; Tuszynski, J.; Kovalenko, A. *J. Chem. Theory Comput.* **2010**, *6*, 607.

Protecting High Energy Barriers: A New Equation to Regulate Boost Energy in Accelerated Molecular Dynamics Simulations

William Sinko,^{*,†,||} César Augusto F. de Oliveira,^{*,‡,§,||} Levi C. T. Pierce,[‡] and J. Andrew McCammon^{†,‡,§}

[†]Biomedical Sciences Program, University of California San Diego, La Jolla, California 92093-0365, United States

[‡]Department of Chemistry & Biochemistry, Department of Pharmacology, and NSF Center for Theoretical Biological Physics, University of California San Diego, La Jolla, California 92093-0365, United States

[§]Howard Hughes Medical Institute, University of California San Diego, La Jolla, California 92093-0365, United States

S Supporting Information

ABSTRACT: Molecular dynamics (MD) is one of the most common tools in computational chemistry. Recently, our group has employed accelerated molecular dynamics (aMD) to improve the conformational sampling over conventional molecular dynamics techniques. In the original aMD implementation, sampling is greatly improved by raising energy wells below a predefined energy level. Recently, our group presented an alternative aMD implementation where simulations are accelerated by lowering energy barriers of the potential energy surface. When coupled with thermodynamic integration simulations, this implementation showed very promising results. However, when applied to large systems, such as proteins, the simulation tends to be biased to high energy regions of the potential landscape. The reason for this behavior lies in the boost equation used since the highest energy barriers are dramatically more affected than the lower ones. To address this issue, in this work, we present a new boost equation that prevents oversampling of unfavorable high energy conformational states. The new boost potential provides not only better recovery of statistics throughout the simulation but also enhanced sampling of statistically relevant regions in explicit solvent MD simulations.

INTRODUCTION

Molecular dynamics simulation (MD) is one of the most common tools used by computational chemists to study the dynamic behavior of biomolecules.^{1,2} However, conventional MD techniques (cMD) are still limited to relatively short time scales, which hinder observation of conformational transitions that are essential for protein function.^{1,3} Most of these transitions occur on a time scale of milliseconds to seconds or longer and often involve the rare crossing of high energy barriers. In an effort to extend the time scale of all-atom molecular dynamics simulations of biomolecules, our group recently proposed an enhanced sampling technique called accelerated molecular dynamics (aMD). This method, which is based on the hyperdynamics technique introduced by Voter,⁴ has been shown to increase conformational sampling of biomolecules over cMD.³ Recently, our group has been successfully using aMD in a wide range of applications and biological systems.^{3,5–11}

Two major implementations of the boost equation for aMD have been proposed. In the original implementation, the boost potential is defined according to eq 1.^{3,5}

$$\Delta V^a = \frac{(E_1 - V(r))^2}{(\alpha_1 + E_1 - V(r))} \quad (1)$$

A continuous non-negative boost potential function, ΔV^a , is added to the original potential surface, $V(r)$, such that regions around the energy minima are raised and those near high barriers or saddle points are left unaffected. Thus, whenever $V(r)$ is below a chosen threshold boost energy, E_1 , the simulation is performed on the modified potential $V^*(r) = V(r) + \Delta V^a$; otherwise, sampling is performed on the original potential $V^*(r) = V(r)$.

The parameter α_1 modulates roughness and the depth of the energy minima on the modified surface.

To recover the correct canonical ensemble, each frame of the simulation must be reweighted using the Boltzmann factor $e^{\beta\Delta V^a[r]}$. Since the lowest energy wells may be associated with the largest boost values, the reweighting can have a detrimental effect on the statistics.^{8,12}

To address this issue, a second implementation was introduced in which energy barriers are modified, instead of energy minima.⁸

$$\Delta V^b = \frac{(V(r) - E_1)^2}{(\alpha_1 + V(r) - E_1)} \quad (2)$$

A continuous negative boost potential function, $\Delta V^b(r)$, is added to the original potential surface, $V(r)$, such that regions around the energy barriers are lowered and those near the minima are left unaffected. Thus, whenever $V(r)$ is above the boost energy, E_1 , the simulation is performed on the modified potential $V^*(r) = V(r) - \Delta V^b$; otherwise, sampling is performed on the original potential $V^*(r) = V(r)$.

This implementation improves the statistical reweighting problem by allowing much of the simulation to remain in the original potential surface, which in this case needs no reweighting. However, application of ΔV^b tends to oversample high energy regions of the potential landscape. As can be seen in eq 2, the boost potential is proportional to the difference $V(r) - E_1$, and as a consequence regions of the potential surface displaying large $V(r)$ (or high-energy regions) are affected significantly more than

Received: April 6, 2011

Published: November 21, 2011

regions with relatively low energy barriers. When applied to large systems, such as proteins, the simulation tends to be biased toward high energy regions of the potential landscape. In small systems, application of ΔV^b revealed promising results when combined with free energy calculations, such as thermodynamic integration (TI).⁸

In this work, we describe a new boost potential (eq 3) in an attempt to combine the strengths of the two previous implementations.

RESULTS AND DISCUSSION

A possible way to overcome the sampling issues associated with ΔV^b is to modify the boost potential equation so that its magnitude reduces significantly at large values of $V(r) - E$.

New equation ΔV^c :

$$\Delta V^c = \frac{(V(r) - E_1)^2}{(\alpha_1 + V(r) - E_1)(1 + e^{-(E_2 - V(r))/\alpha_2})} \quad (3)$$

We defined a second energy level (E_2) in order to return the modified potential surface back to the original one whenever the potential energy of the system is larger than E_2 . This boost equation is shown above as ΔV^c (eq 3). The second energy level allows the user to define a window of acceleration between E_1 and E_2 . To regulate the return to the original potential upon crossing E_2 , a second parameter α is required (α_2). The term in the large brackets in the denominator is responsible for bringing the boost to zero when the potential energy $V(r)$ is higher than E_2 . Thus, when $V(r)$ is higher than E_2 , $(1 + e^{-(E_2 - V(r))/\alpha_2})$ tends to a very large positive number, and as a result, the modified potential converges to the original one, $V(r)$. On the other hand, when $V(r)$ is lower than E_2 , the term $(1 + e^{-(E_2 - V(r))/\alpha_2})$ tends to 1, which results in $\Delta V^c = \Delta V^b$ or eq 2.

We explored the new boost equation by creating a hypothetical one-dimensional potential using the analytical equation below:

$$V(r) = -200 + 50 \times \left(\cos(r \times \pi) - 1 - \frac{r^2}{r + (1-r) \times \sqrt{\frac{3-r}{4}}} \right) \quad (4)$$

Figure 1 displays the effect of boost energy E (E_1 and E_2) and α (α_1 and α_2) on eqs 2 and 3. The upper solid black line represents the unmodified potential $V(r)$, while the lower solid black line represents modified potential $V(r)^*$ generated after the application of eq 2, ΔV^b . Boost energies E are shown as dashed lines. The solid colored lines represent different modified potentials, $V(r)^*$, generated by ΔV^c with different sets of parameters. Figure 1A shows that high energy barriers can be selectively protected by setting different values of E_2 . It is worth noting that the modified potential generated by ΔV^c follows closely along ΔV^b until the difference between E_2 and E_1 is similar to the difference between $V(r)$ and E_1 (Figure 1A and B).

Like in the original implementation, the degree of acceleration is controlled by the parameter α_1 and E_1 . Parameter α_2 controls how strongly energy barriers higher than E_2 are protected. For instance, when V is higher than E_2 , in the limit $\alpha_2 \rightarrow \infty$, the term $(1 + e^{-(E_2 - V(r))/\alpha_2}) \rightarrow 2$ and ΔV^c converges to $1/2\Delta V^b$, and

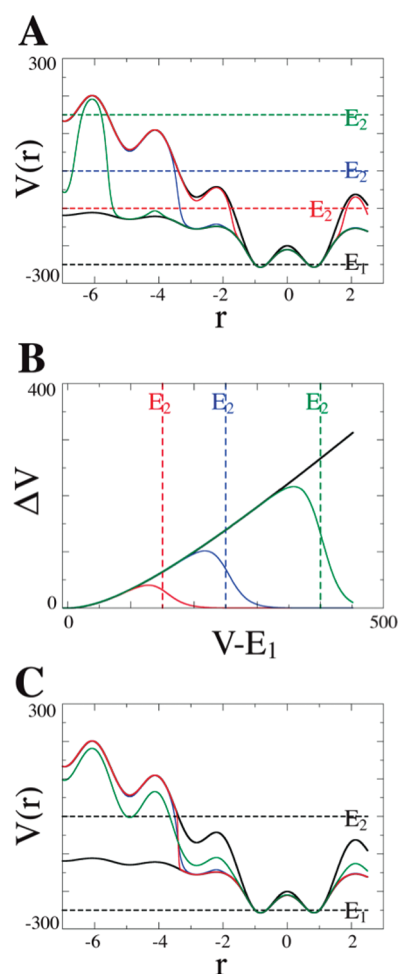


Figure 1. Hypothetical one-dimensional potential representing the effect of ΔV^c . In all charts, $\alpha_1 = 200$ and $E_1 = -250$. The upper and lower solid black lines represent the original potential and the modified potential generated with ΔV^b , respectively. This color scheme is used throughout. (A) Effects of different parameters E_2 (dashed colored lines) on the modified potential generated with ΔV^c (solid colored lines). (B) Boost levels ΔV^b (solid black line) and ΔV^c (colored lines) as $V(r)$ moves away from E_1 . For both A and B, $\alpha_2 = 15$ and $E_2 = -100$ (red), 0 (blue), and 150 (green). (C) Effect of varying α_2 parameter on ΔV^c : $\alpha_2 = 3$ (red), 15 (blue), and 75 (green) with $E_2 = 0$.

as a result, large energy barriers are not effectively protected anymore. On the other hand, when $\alpha_2 \rightarrow 0$, the term $(1 + e^{-(E_2 - V(r))/\alpha_2}) \rightarrow \infty$ and $\Delta V^c \rightarrow 0$, thus keeping all energy regions, where $V(r)$ is higher than E_2 , unchanged. Figure 1C displays the effects of α_2 on $V(r)^*$.

Although this new implementation introduces two new parameters, E_2 and α_2 are easily estimated. Initial guesses for α_2 are based on the hypothetical one-dimensional potential shown in Figure 1. To keep the underlying shape of the original potential surface and effectively protect energy barriers higher than E_2 , α_2 is recommended to be proportional to the difference $\sim (E_2 - E_1)$. More specifically, we estimate α_2 to be between 20 and 60% of the difference $(E_2 - E_1)$. Energy levels E_1 and E_2 are estimated from short cMD simulations. Since ΔV^c is only effectively applied to the system whenever the potential $V(r)$ is higher than E_1 , it is important to not set E_1 much higher than the average potential energy of system, $\langle V(r) \rangle$, in order to guarantee a minimum degree of acceleration. In this work, $V(r)$ and $\langle V(r) \rangle$ correspond

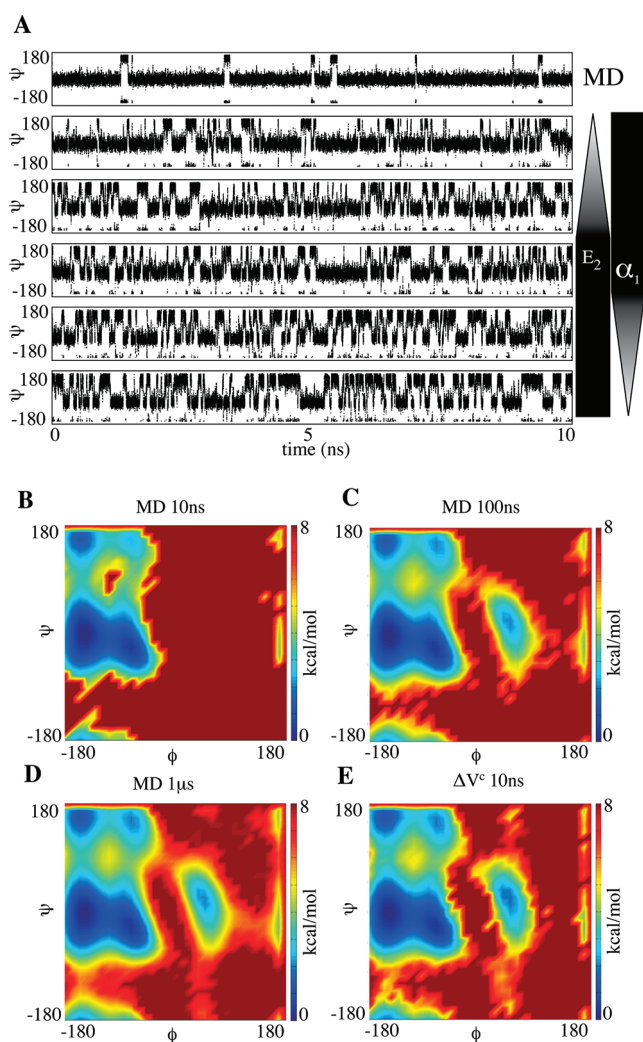


Figure 2. Alanine dipeptide simulation results. (A) Ψ angle values obtained from cMD and five different aMD simulations. From top to bottom, aMD parameters were set to $E_2 = E_1 + 15$ and $\alpha_1 = 5$, $E_2 = E_1 + 20$ and $\alpha_1 = 5$, $E_2 = E_1 + 25$ and $\alpha_1 = 5$, $E_2 = E_1 + 25$ and $\alpha_1 = 2.5$, and $E_2 = E_1 + 25$ and $\alpha_1 = 1.25$. In all simulations, E_1 and α_2 were set to 10 and 5, respectively. Weighted free energy density plots obtained from cMD (B, C, D) and aMD with ΔV^c (E). All values are in kcal/mol.

to the instantaneous and average dihedral energy, respectively. E_2 is simply defined as $E_2 = E_1 + \Delta E$, where ΔE is the highest energy barrier that is allowed to be crossed. The selection of optimum boost parameters is bound to be system dependent. For this reason, short aMD runs are strongly recommended to fine-tune parameters α_1 and E_1 . Failure in obtaining suitable parameters may lead to two possible scenarios: (i) No or extremely low acceleration is effectively applied to the system. In this case, aMD and cMD will likely generate very similar trajectories. (ii) Extremely high acceleration is applied to system, which results in serious structural and energetic instabilities.

Unless otherwise stated, all simulations were performed applying the boost potential ΔV^c to the dihedral terms of the potential energy function. Enhanced sampling techniques, such as aMD, based on the dihedral energy contributions have been successfully used to effectively enhance conformational sampling of biomolecules.^{10,13–17} The approach presented in this work can be easily extended to the nonbonded energy terms via the dual boost

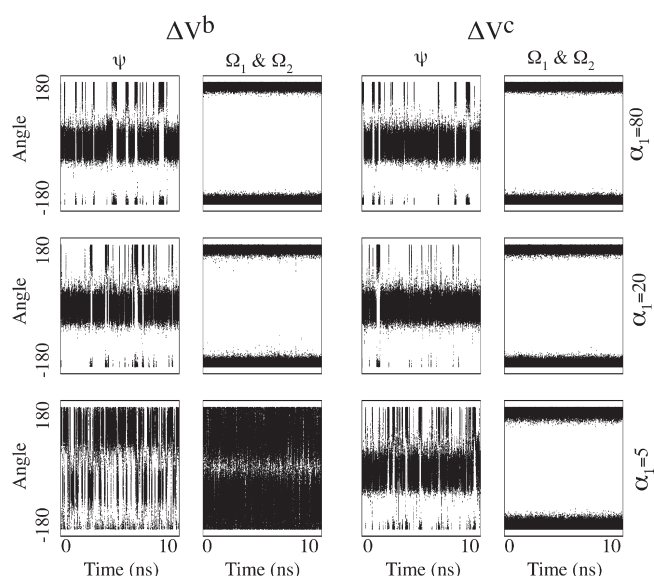


Figure 3. ψ and Ω angle values obtained from aMD simulations with boost potentials ΔV^b and ΔV^c . In all simulations, $E_1 = 10.0$ and α_1 were set as shown on the far right. Additional parameters for aMD with ΔV^c were set to $E_2 = E_1 + 15$ and $\alpha_2 = 5$. All values are in kcal/mol.

method.⁹ To investigate the use of the new boost equation ΔV^c , we first compared our aMD simulations results of fully solvated alanine dipeptide to cMD protocols. Alanine dipeptide has been extensively studied as a model system to evaluate free energy and conformational change in biomolecular simulations.^{18–24} Figure 2A displays the time evolution of the Ψ angle during the cMD and five aMD simulations of 10 ns. As can be clearly seen, the number of Ψ transitions dramatically increases as we modify boost parameters E_2 and α_2 . Figure 2B–E show the free-energy density plots obtained from cMD simulations of 10 ns, 100 ns, and 1 μ s and an aMD simulation of 10 ns. The free-energy density plots were calculated from the population of states sampled on each simulation. To recover the corrected canonical ensemble, each frame of the aMD trajectory was Boltzmann weighted by its respective boost factor. Figure 2B reveals that the conformational sampling obtained from 10 ns of cMD is mainly restricted to α -helical ($\Phi < 0^\circ$ and $-60^\circ < \Psi < 0^\circ$) and β -strand regions ($\Phi < 0^\circ$ and $120^\circ < \Psi < 180^\circ$), with the α -helical region displaying the most populated states. A significant increase in conformational sampling is evident when the cMD is extended to 100 ns (Figure 2C). The most pronounced change can be seen in the left-handed α -helix region ($\Phi \sim 50$ and $\Psi \sim 50$), which is now well sampled and is not observed in the cMD of 10 ns. A dramatic increase in the number of transitions between the α -helical and β -strand regions is also noted. To provide some insights concerning the time scale accessed by our aMD runs, we further extended the cMD simulation to 1 μ s. A comparison of Figure 2D and E clearly shows that there is good agreement between the regions sampled by our short aMD of 10 ns and the cMD of 100 ns and 1 μ s. For the alanine dipeptide system, these results suggest that aMD simulations with ΔV^c can accelerate conformational sampling by at least 10–100 fold.

While boosting through energy barriers is important for sampling, limiting the boost to reduce the population of thermodynamically unfavorable states is equally important. To illustrate the advantage of ΔV^c and its boost limiting capabilities over ΔV^b , we analyzed and compared the Ψ and Ω angle transitions

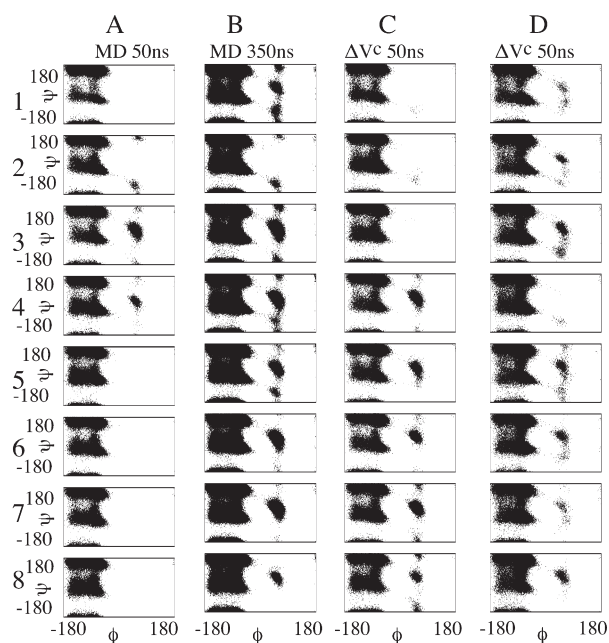


Figure 4. Decalanine Φ – Ψ angles distribution obtained from cMD and aMD simulations. For the aMD simulations with ΔV^c parameters were set to $E_1 = 74$, $E_2 = E_1 + 25$, $\alpha_2 = 5$, $\alpha_1 = 30$ (C), and $\alpha_1 = 15$ (D). All values are in kcal/mol.

(cis/trans isomerization) obtained from the alanine dipeptide simulations in both implementations. As seen in Figure 3, as the degree of acceleration is increased (by reducing the value of parameter α_1), ΔV^b dramatically increases not only Ψ but also Ω dihedral transitions. Conversely, ΔV^c promotes a very similar increase in Ψ dihedral transitions without affecting the Ω dihedral angles. This result confirms the capability of ΔV^c to accelerate conformational transitions by selectively crossing energy barriers lower than the predefined energy level. It is worth mentioning that ΔV^b notably undersamples the normally preferred region $-50 > \Psi > +50$ under high acceleration conditions.

To evaluate the applicability of equation ΔV^c to biomolecules, we also performed aMD studies on a more complex model system, decalanine.²⁵ Figure 4 displays the distribution of eight Φ – Ψ angles monitored along two cMD simulations of 50 ns and 350 ns and two independent aMD of 50 ns. All simulations started from a fully solvated and extended conformation. As expected, there is a substantial improvement in conformational sampling when the cMD simulation is extended from 50 ns to 350 ns (Figure 4A and B). Similar results are obtained for Φ – Ψ angles 4–8 when we compare aMD with both cMD simulations (Figure 4A, B, and C). Interestingly, the opposite behavior is observed for Φ – Ψ angles 1–3 (Figure 4C). We attribute this result to the low degree of acceleration used on the aMD simulation. Even though application of ΔV^c enhances conformational transitions of decalanine, the small boost used in this simulation, as a test case, may not generate the 7-fold acceleration expected from Figure 4C and B. To investigate this issue and further explore the capability of ΔV^c , we carried out two extra aMD simulations of 50 ns in which we (a) increased the acceleration by reducing the α_1 value by a factor of 2 (result is shown in Figure 4D) and (b) increased the degree of acceleration by raising the energy level E_2 ($E_2 = E_1 + 35$ kcal/mol), in addition to reducing α_1 by a factor of 2 (Figure S1, Supporting Information). As expected, the different aMD simulations of

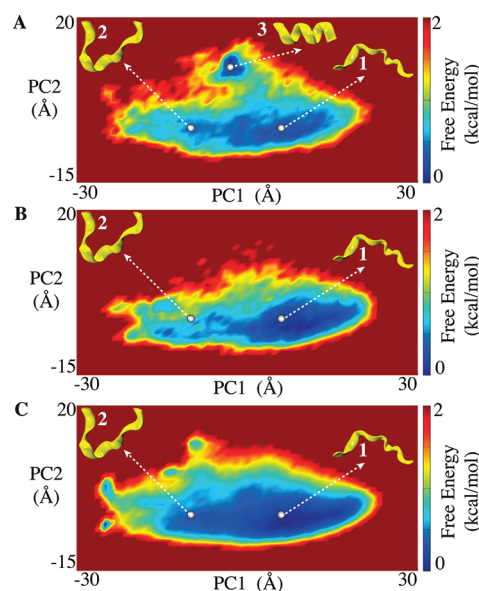


Figure 5. Principle component analysis obtained from decalanine MD simulations. (A) 50 ns of aMD simulation with ΔV^c . Parameters were set to $E_1 = 74$, $E_2 = E_1 + 25$, $\alpha_2 = 5$, and $\alpha_1 = 15$, same as in Figure 4D. (B) 50 ns of cMD simulation and (C) 350 ns of cMD simulation. Structures 1, 2, and 3 shown in yellow represent relevant populated states in PC subspace sampled by aMD and cMD.

50 ns each (Figure 4C,D and Figure S1) cover different regions of the Φ – Ψ subspace, with the more accelerated ones (Figure 4D) showing better agreement with the cMD simulation of 350 ns (Figure 4B). These results also agree with the fact that, by lowering energy barriers, aMD increases the rate of escape from minimum wells and thus generates more diverse trajectories for complex systems with multidimensional energy landscapes such as decalanine. Figure S1 displays the Φ – Ψ angle distributions obtained with the highest degree of acceleration tested. It is worth noting that there is better agreement with the conformational sampling obtained from the 350 ns of cMD, as a result of the longer time scale accessed by this aMD simulation.

Decalanine can adopt numerous secondary structures making it a challenging test case for enhanced sampling methods.²⁵ Principal component analysis (PCA) shows that our ΔV^c aMD simulation explores energy wells that are not adequately sampled by 350 ns of cMD simulation (Figure 5A, B, C). One of these regions represents the state in which decalanine adopts an α -helical conformation, energetically the most stable configuration.²⁵ This folding event is evident in the aMD simulations, but not in the cMD simulations despite the latter being 7-fold longer (Figure 2S).

Free energy calculations are useful in the optimization of compounds for biological targets and host systems.²⁶ However, these calculations usually require a computationally expensive ensemble generation either from Monte Carlo calculations or MD simulations.^{27,28} As previously shown, coupling of aMD methods with free energy calculations, such as thermodynamic integration (TI), revealed promising results when applied to simple model systems.⁸ To further extend the applicability of aMD-based approaches to free energy calculations, in this work, we modified our original implementation by incorporating the boost equation ΔV^c into the TI simulations. As a test case, we calculated the relative free energy difference between Ac_{2-L}-Lys_D-Ala_D-Ala and Ac_{2-L}-Lys_D-Ala_D-Lac bound to vancomycin. This mutation, Ala

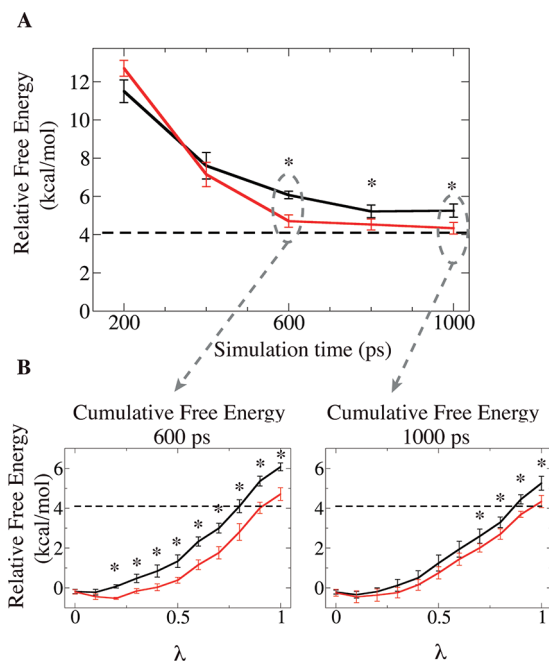


Figure 6. (A) Relative free energy of binding between Ac_{2-L}Lys-D-Ala_{-D}Ala and Ac_{2-L}Lys-D-Ala_{-D}Lac to vancomycin calculated from cMD (solid black line) and aMD with ΔV^c (solid red line). A dashed line displays the experimental value, 4.1 kcal/mol. (B) Cumulative free energy curves calculated from simulations of 600 ps (left) and 1000 ps (right) per λ point. The * shows points where there is no overlapping between error bars.

to Lac, confers to bacteria a resistance against vancomycin.²⁹ The experimental change in binding free energy has been determined to be 4.1 kcal/mol, which corresponds to an approximately 1000-fold decrease in affinity from _DAla to _DLac.³⁰

Figure 6A compares the relative free energy of binding ($\Delta\Delta G$) calculated from TI simulations using cMD and aMD with ΔV^c . To calculate the final free energy values, we divided the trajectories in blocks of 200 ps, with the last block representing the production phase. For example, a TI simulation of 800 ps corresponds to an equilibration phase of 600 ps (three blocks) followed by a collecting data phase of 200 ps, and a TI simulation of 1000 ps corresponds to an equilibration phase of 800 ps (four blocks) followed by the collecting data phase of 200 ps. Thus, the points displayed in Figure 6A reveals how the calculated $\Delta\Delta G$ changes as a function of the equilibration time.

It is worth mentioning that application of ΔV^c notably improves the convergence of $\Delta\Delta G$ when compared to standard cMD TI simulation. In addition, the final free energy value obtained with ΔV^c (4.3 ± 0.3 kcal/mol) shows very good agreement with the experimental value of 4.1 kcal/mol,³⁰ while the final free energy value from TI with cMD is 5.3 ± 0.3 kcal/mol. Since the same force field and simulation conditions were applied to both TI simulations, we attribute this difference solely on the conformational sampling enhancement provided by the ΔV^c . Moreover, the error associated with each point suggests that the faster convergence toward the final free energy value is statistically relevant. Interestingly, the cumulative free energy values (Figure 6B) demonstrate that the TI simulations coupled with cMD are indeed converging toward the ones coupled with aMD as we increase the simulation time. Hence, inaccuracies in the final value are likely to be primarily due to the lack of convergence

on λ points. These results indicate that ΔV^c can effectively enhance conformational sampling when coupled with TI simulations and hence shorten the equilibration period required for accurate free energy calculation.

COMPUTATIONAL METHODS

ΔV^c was implemented in the AMBER10 code³¹ as previously reported.⁸

$$V^*(r) = V(r) - \Delta V^c$$

$$\Delta V^c = \begin{cases} \frac{(V(r) - E_1)^2}{(\alpha_1 + V(r) - E_1)(1 + e^{-(E_2 - V(r))/\alpha_2})} & V(r) > E_1 \\ 0 & V(r) \leq E_1 \end{cases} \quad (5)$$

All cMD, aMD and TI simulations were performed using a modified version of the sander module of the AMBER10 package.³¹ TIP3P water molecules were used to solvate both the alanine dipeptide and decalanine systems.³² A buffer region of 10 or 12 Å was used in all systems. To eliminate any steric clashes, 100 steps of conjugate gradient minimization were performed on all systems. To bring the systems to the right density, we carried out cMD simulations of 50 ps in which the NPT ensemble was applied. Then, long cMD and aMD simulations were performed in which the NVT ensemble was applied. All bonds involving hydrogen atoms were constrained using the SHAKE algorithm.³³ The temperature and pressure were controlled using weak coupling to external temperature and pressure baths.³⁴ Electrostatic interactions were computed via PME (particle mesh Ewald summation) with a cutoff of 8.0 Å. All simulations were performed at temperature of 300 K. In all accelerated simulations, the boost potential was based on the dihedral energy. Principal components analysis was performed using the ptraj module of the AMBER10 package. All cMD simulations were projected onto the PC subspace obtained from the aMD simulation displayed at Figure 4D. Alignment of the trajectory was performed on backbone atoms of decalanine.

To study the use of the new boost equation on thermodynamic integration calculations, we calculated the relative difference in the free energy of binding of Ac_{2-L}Lys-D-Ala_{-D}Ala and Ac_{2-L}Lys-D-Ala_{-D}Lac to a vancomycin dimer, starting from the crystal structure of Ac_{2-L}Lys-D-Ala_{-D}Ala bound to vancomycin (PDB ID: 1FVM). The glycopeptides and vancomycin were parametrized using Antechamber. The system was solvated in a cubic box of TIP3P water molecules, with a buffer region of 10 Å.³¹ Owing to the strong correlation between glycopeptides binding affinity and vancomycin dimerization,³⁵ we simulated the “back to back” homodimer of vancomycin, as present in the X-ray crystal structure. Both ligands were included in the model and modified alchemically.

TI simulations were performed with nine equally spaced λ parameters ($\lambda = 0.1$ to 0.9) in solution and in the vancomycin receptor. In all transformations, electrostatic and van der Waals contributions were decoupled and computed separately. More specifically, in this work, the alchemical transformation of _DAla to _DLac was carried out in three steps: (i) removal of partial charges of the NH group from _DAla, (ii) transformation of van der Waals parameters of the NH group to the O (oxygen) atom, and (iii) partial charge creation on the O (oxygen) atom. Softcore potentials were used for step ii.^{36,37} The $\Delta V/\Delta\lambda$ values were calculated over a production period of 200 ps along with five equilibration periods 0, 200, 400, 600, and 800 ps. The final

free energy values were averaged over three independent (with reassigned initial atomic velocities) cMD or aMD simulations. As previously shown, in order to recover the correct canonical ensemble, $\Delta V/\Delta\lambda$ values collected from aMD runs were reweighted by their respective boost factor $e^{\beta\Delta V[r]}$.^{3,8}

Error bars were calculated using, $\sigma_{\langle A \rangle} \approx \sigma/\sqrt{M}$ where M is the number of independent simulations and $\sigma_{\langle A \rangle}$ is the standard deviation of the average value A obtained from M independent data values ($M = 3$ in all cases). An analysis of the trajectories was performed using ptraj.³¹

Our aMD parameters were estimated on the basis of the average dihedral energy term obtained from short cMD simulations. For all alanine dipeptide aMD simulations, parameter E_1 was set to 10 kcal/mol. Parameter α_2 was set to 5 kcal/mol, which corresponds to 0.2 to 0.33($E_2 - E_1$). In Figure 2A, from top to bottom, aMD simulations used the following parameters: $E_2 = E_1 + 15$ and $\alpha_1 = 5$, $E_2 = E_1 + 20$ and $\alpha_1 = 5$, $E_2 = E_1 + 25$ and $\alpha_1 = 5$, $E_2 = E_1 + 25$ and $\alpha_1 = 2.5$, and $E_2 = E_1 + 25$ and $\alpha_1 = 1.25$. In Figure 2E, the aMD simulation used the parameters $E_2 = E_1 + 15$ and $\alpha_1 = 5$. In Figure 3, $E_2 = E_1 + 15$ (for ΔV^c) and α_1 were varied as indicated in the far right column.

Boost parameters for decalanine simulations were $E_1 = 74$, $E_2 = E_1 + 25$, $\alpha_1 = 30$, and $\alpha_2 = 5$. Boost parameters for the vancomycin-glycopeptides simulations were $E_1 = 211$, $E_2 = E_1 + 25$, $\alpha_1 = 30$, and $\alpha_2 = 15$.

CONCLUSION

In this work, we introduced a new boost equation, ΔV^c , for aMD simulations aiming to overcome sampling issues introduced by ΔV^b . Since energy barriers located above a predefined energy level can now be protected, the new boost equation ΔV^c provided much better control over high energy regions of the conformational landscape when compared to ΔV^b . We used two model systems, alanine dipeptide and decalanine, to study the applicability and efficiency of ΔV^c in enhancing conformational sampling. In both cases, the new boost potential not only provides better recovery of statistics throughout the simulation but also enhanced sampling of statistically relevant regions in explicit solvent MD simulations. When coupled with thermodynamic integration, our results indicate that ΔV^c can effectively enhance conformational sampling and accelerate convergence for a more accurate free energy calculation.

ASSOCIATED CONTENT

Supporting Information. Additional simulation and analysis of decalanine including the Φ – Ψ angles distribution for an aMD simulation and calculation of the RMS deviation from the α -helix conformation. This information is available free of charge via the Internet at <http://pubs.acs.org/>.

AUTHOR INFORMATION

Corresponding Author

*Phone: 858-822-1083. Fax: 858-534-4974. E-mail: wsinko@ucsd.edu, cesar@mccammon.ucsd.edu.

Author Contributions

^{||}These authors contributed equally.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

This work was supported by the Molecular Biophysics Training Grant GM08326 (WS), the National Science Foundation Grant MCB-0506593, NBCR, CTBP, Howard Hughes Medical Institute (JAM), and National Institutes of Health Grant GM31749 (JAM).

REFERENCES

- (1) Karplus, M.; McCammon, J. A. *Nat. Struct. Biol.* **2002**, *9*, 646.
- (2) Adcock, S. A.; McCammon, J. A. *Chem. Rev.* **2006**, *106*, 1589.
- (3) Hamelberg, D.; Mongan, J.; McCammon, J. A. *J. Chem. Phys.* **2004**, *120*, 11919.
- (4) Voter, A. F. *Phys. Rev. Lett.* **1997**, *78*, 3908.
- (5) Markwick, P. R.; Cervantes, C. F.; Abel, B. L.; Komives, E. A.; Blackledge, M.; McCammon, J. A. *J. Am. Chem. Soc.* **2010**, *132*, 1220.
- (6) de Oliveira, C. A.; Hamelberg, D.; McCammon, J. A. *J. Phys. Chem. B* **2006**, *110*, 22695.
- (7) de Oliveira, C. A.; Hamelberg, D.; McCammon, J. A. *J. Chem. Phys.* **2007**, *127*, 175105.
- (8) de Oliveira, C. A.; Hamelberg, D.; McCammon, J. A. *J. Chem. Theory Comput.* **2008**, *4*, 1516.
- (9) Hamelberg, D.; de Oliveira, C. A.; McCammon, J. A. *J. Chem. Phys.* **2007**, *127*, 155102.
- (10) Hamelberg, D.; McCammon, J. A. *J. Am. Chem. Soc.* **2005**, *127*, 13778.
- (11) Hamelberg, D.; Shen, T.; Andrew McCammon, J. *J. Chem. Phys.* **2005**, *122*, 241103.
- (12) Shen, T.; Hamelberg, D. *J. Chem. Phys.* **2008**, *129*, 034103.
- (13) Hamelberg, D.; Shen, T.; McCammon, J. A. *J. Am. Chem. Soc.* **2005**, *127*, 1969.
- (14) Shen, T.; Hamelberg, D.; McCammon, J. A. *Phys. Rev. E* **2006**, *73*, 041908.
- (15) Markwick, P. R. L.; Bouvignies, G.; Blackledge, M. *J. Am. Chem. Soc.* **2007**, *129*, 4724.
- (16) Yang, W.; Li, H.; Min, D.; Liu, Y. *J. Chem. Phys.* **2007**, *127*.
- (17) Yang, W.; Zheng, L. Q. *J. Chem. Phys.* **2008**, *129*.
- (18) Yonezawa, Y.; Fukuda, I.; Kamiya, N.; Shimoyama, H.; Nakamura, H. *J. Chem. Theory Comput.* **2011**, *7*, 1484.
- (19) Ng, K. M.; Solayappan, M.; Poh, K. L. *Comput. Biol. Chem.* **2011**, *35*, 19.
- (20) Ferguson, A. F.; Panagiotopoulos, A. Z.; Debenedetti, P. G.; Kevrekidis, I. G. *J. Chem. Phys.* **2011**, *134*.
- (21) Cruz, V.; Ramos, J.; Martinez-Salazar, J. *J. Phys. Chem. B* **2011**, *115*, 4880.
- (22) Vondrasek, J.; Vymetal, J. *J. Phys. Chem. B* **2010**, *114*, 5632.
- (23) Ishizuka, R.; Huber, G. A.; McCammon, J. A. *J. Phys. Chem. Lett.* **2010**, *1*, 2279.
- (24) Adams, J. P.; Smith, D. A. *Abstr. Pap.—Am. Chem. Soc.* **1993**, *206*, 42.
- (25) Hénin, J.; Fiorin, G.; Chipot, C.; Klein, M. L. *J. Chem. Theory Comput.* **2009**, *6*, 35.
- (26) Michel, J.; Foloppe, N.; Essex, J. W. *Mol. Inf.* **2010**, *29*, 570.
- (27) Gilson, M. K.; Moghaddam, S.; Inoue, Y. *J. Am. Chem. Soc.* **2009**, *131*, 4012.
- (28) Kollman, P. *Chem. Rev.* **1993**, *93*, 2395.
- (29) Bugg, T. D. H.; Wright, G. D.; Dutka-Malen, S.; Arthur, M.; Courvalin, P.; Walsh, C. T. *Biochemistry* **1991**, *30*, 10408.
- (30) McComas, C. C.; Crowley, B. M.; Boger, D. L. *J. Am. Chem. Soc.* **2003**, *125*, 9314.
- (31) Case, D. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Zhang, W.; Merz, K. M.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, A. P. A. *AMBER10*; University of California: San Francisco, CA, 2008.

- (32) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.
- (33) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comp. Phys.* **1977**, *23*, 327.
- (34) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684.
- (35) Mackay, J. P.; Gerhard, U.; Beuregard, D. A.; Williams, D. H.; Westwell, M. S.; Searle, M. S. *J. Am. Chem. Soc.* **1994**, *116*, 4581.
- (36) Zacharias, M.; Straatsma, T. P.; McCammon, J. A. *J. Chem. Phys.* **1994**, *100*, 9025.
- (37) Beutler, T. C.; Mark, A. E.; van Schaik, R. C.; Gerber, P. R.; van Gunsteren, W. F. *Chem. Phys. Lett.* **1994**, *222*, 529.

Kepler Predictor—Corrector Algorithm: Scattering Dynamics with One-Over-R Singular Potentials

Andreas Markmann,^{*,†} Frank Graziani,[‡] and Victor S. Batista^{*,†}

[†]Department of Chemistry, Yale University, P.O. Box 208107, New Haven, Connecticut 06520-8107, United States

[‡]Center for Applied Scientific Computing and B-Division, Lawrence Livermore National Laboratory, P.O. Box 808, Livermore, California, United States

ABSTRACT: An accurate and efficient algorithm for dynamics simulations of particles with attractive $1/r$ singular potentials is introduced. The method is applied to semiclassical dynamics simulations of electron–proton scattering processes in the Wigner-transform time-dependent picture, showing excellent agreement with full quantum dynamics calculations. Rather than avoiding the singularity problem by using a pseudopotential, the algorithm predicts the outcome of close-encounter two-body collisions for the true $1/r$ potential by solving the Kepler problem analytically and corrects the trajectory for multiscattering with other particles in the system by using standard numerical techniques (e.g., velocity Verlet, or Gear Predictor corrector algorithms). The resulting integration is time-reversal symmetric and can be applied to the general multibody dynamics problem featuring close encounters as occur in electron–ion scattering events, in particle–antiparticle dynamics, as well as in classical simulations of charged interstellar gas dynamics and gravitational celestial mechanics.

1. INTRODUCTION

Understanding the dynamics of particles mutually attracted by $1/r$ singular potentials is a problem common to a wide range of systems in chemistry, biology, physics, and astronomy, including classical and semiclassical studies of electron transfer,¹ excess electrons in liquids,^{2,3} ionic states,⁴ electron scattering and trapping in ionic liquids or solids,^{5–7} attractive plasmas with nuclei and electrons,^{8,9} particle–antiparticle dynamics,¹⁰ dynamics of charged interstellar gas particles,¹¹ and celestial mechanics.^{12,13} However, serious numerical problems typically arise in classical and semiclassical simulations when particles gravitate into each other and the potential gradients (or accelerations) diverge. To avoid this type of Coulomb (or gravitational) catastrophe problem, simulation studies often rely on pseudopotentials where the essential singularities of the potentials are artificially removed. Such approximations lead to integration methods that are both practical and useful for simulations of scattering events with large impact parameters, as typically observed in low-energy collisions of particles with repulsive cores. However, close-encounter collisions are beyond the capabilities of pseudopotential methods, and more rigorous methodologies have to be employed. This paper introduces a predictor–corrector algorithm for dynamics simulations of particles evolving on attractive $1/r$ singular potentials. The method is rigorous and efficient, even when modeling close-encounter collisions. Its application to semiclassical dynamics simulations of electron–proton scattering in the Wigner-transform time-dependent picture shows excellent agreement with full quantum dynamics calculations.

The Coulomb catastrophe problem could be avoided by using a quantum treatment of the attractive interaction, setting a lower limit for the bound state in the potential. However, quantum dynamics methods are computationally demanding and scale poorly (i.e., exponentially) with the number of strongly coupled

particles in the system. For example, quantum dynamics simulations of dense proton plasmas with electrons are usually computationally impractical. Accurate simulations of such systems thus require a rigorous solution to the Coulomb catastrophe problem within the framework of linear scaling particle simulation techniques, analogous to standard molecular dynamics simulations. When implemented with adaptive time-step integrators, such methods are capable of accurately simulating point particles interacting through singular attractive pair potentials, bypassing energy conservation problems associated with divergent accelerations.

Adaptive time-step integrators are the most common techniques applied to ensure energy conservation for systems with rapidly changing potential gradients (or accelerations). However, such methods are hopeless for simulations of close encounters since the integration time-steps converge to zero as the accelerations diverge and bring the simulation to a halt. [Specifically, the time step becomes so small that additive terms that include the time step as a factor become smaller than the least significant digit of the coordinate, so coordinates no longer evolve.] This happens even if individual time-steps are used for each particle.^{12,14} Conversely, imposing a minimal time-step yields trajectories that eventually violate energy conservation, prompting the common practice of using pseudopotentials to artificially remove the singularities and ensure energy conservation.

Smoothed (pseudo)potentials with a finite value at the origin have been postulated for electron dynamics^{5,8} as well as for ion dynamics¹⁵ and astrophysical simulations.^{12,16} The Coulomb potential is also commonly switched-off at small interparticle distances through the use of switching functions (e.g., the error function).^{7,9} These cases are typically distinguished from the

Received: June 29, 2011

Published: November 08, 2011

construction of pseudopotentials based on physical insights, such as the screening effect of electrons in the conduction band of metals.⁶ In ionic solids (e.g., alkali halides, oxide insulators, or semiconductors), however, there are no electrons in the conduction band that could offer screening of Coulombic interactions. Therefore, alternative methods are required.

Changes in the potential usually alter the underlying dynamics of the systems yielding artificial effects that disappear only when the pseudopotentials become more and more similar to the true potentials. In that limit, however, the numerical problems due to large gradients usually reappear. An approach that avoids changing the singular potential has been developed for gravitational systems, implementing a change of variables that regularizes the dynamics (e.g., the Kustaanheimo–Stiefel (KS) regularization and related methods^{12,17–21}) and applies standard numerical integration for the new variables rather than for the original coordinates. This requires transformation of the time variable and coordinates that depend on the interparticle distance for the pair of particles experiencing a close encounter. Great care must be taken to keep track of transformations for multiple close encounters and to match up time-steps so that the interaction with other particles present in the simulation is properly accounted for, while not sacrificing efficiency.^{20,21} As a consequence of these complications, the method is, to our knowledge, implemented in only a few stellar codes.²²

In this paper, we introduce a simple Kepler predictor corrector (KPC) algorithm where close-encounter collisions are integrated analytically by solving the Kepler two-body problem without altering (or smoothing) the Coulomb potential and updating coordinates based on the residual potential due to particles not participating in the close encounter. A simple well-known example for such an approach is the lightly damped harmonic oscillator, where the frequency of the resulting oscillation is approximately the same as that of the underlying undamped oscillator and only the amplitude may be viewed as modulated.²³ In molecular dynamics, the analytic solution of the harmonic oscillator has been employed by splitting linear molecule Hamiltonians into a harmonic and anharmonic part and treating high frequency components of the molecular vibrations analytically, while treating low frequency components numerically.^{24–29} In biological systems, a large speedup was achieved by treating water molecules as rigid and using analytic solutions for their motion.³⁰

For the $1/r$ problem, an update scheme based on a single momentum shift has been proposed previously for the purpose of simulating ion collisions.³¹ We propose a method that corrects the predicted scattering trajectories with an additional term rigorously derived from numerical integrators (e.g., velocity Verlet or Gear Predictor Corrector algorithms), accounting for the regular influence of other particles (or external fields) in the system. The integration method as a whole is time-reversal symmetric, does not require any time-variable transformation, and can be applied to the general multibody dynamics problem with close encounters, as in electron–proton scattering processes, particle–antiparticle dynamics, and charged gas dynamics. Algorithmically, the complete potential is separated into two parts, including an integrable two-body term that is dominant at close encounters and a correction due to interactions with all other particles (or potentials) in the system. With an integrable two-body problem (as is the case of Coulombic $1/r$ or van der Waals $1/r^6$ potentials),³² the analytic solution of close encounters is employed, and the effect due to interactions with all other particles is introduced by augmenting the analytic solution with

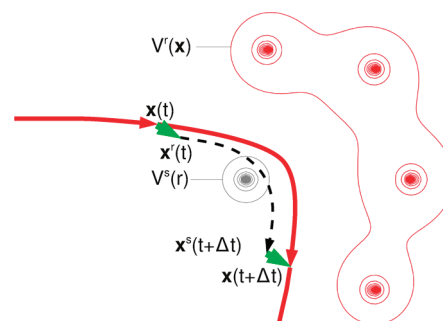


Figure 1. KPC trajectory (red line) of a particle scattered by a singularity $V^s(r)$, obtained by solving the Kepler problem (dashed line), taking into account the effect of a weakly varying residual potential $V^r(x)$ (bold green lines) as described in the text.

additional terms that stem from the numerical integrator (e.g., the velocity Verlet method^{33,34}). The resulting KPC method allows for integration time-steps on the same order as used for regular potentials and hence reduces the numerical effort for problems with close encounters. It is applicable for any integrable singular pair potential,³² including the family of integrable spherically symmetric singular pair potentials that are the focus of this paper.

The paper is organized as follows. Section 2 introduces the KPC method as applied to modeling the dynamics of a particle evolving on a potential with multiple singularities and its generalization to multibody dynamics. Section 3 describes its implementation for semiclassical simulations of electron–proton scattering in the Wigner-transform time-dependent picture. Section 4 presents concluding remarks in perspective of the KPC limitations and applicability as a general method.

2. KEPLER PREDICTOR CORRECTOR ALGORITHM

2.1. Single Particle Collision with Multiple Singularities.

Consider the general case of a single particle moving in a stationary potential with multiple singularities, as shown in Figure 1. The total potential V acting on the particle is composed of the spherically symmetric potential V^s with the singularity nearest to the particle, and the residual potential V^r :

$$V(\mathbf{x}) = V^s(r) + V^r(\mathbf{x}) \quad (1)$$

As an example, we consider the semiclassical trajectory of a fast electron undergoing multiple scattering through proton plasma, under the Born–Oppenheimer approximation, where $V(\mathbf{x})$ is defined as the Coulomb potential due to the closest proton $V^s(\mathbf{x})$ plus the sum of Coulomb potentials $V^r(\mathbf{x})$ due to the other protons in the plasma.

During a close encounter with the singularity $V^s(r)$, the electron scattering force is dominated by this closest proton, and accurate numerical integration faces several difficulties, including the following:

1. the large absolute value of potential energy,
2. the very large norm of the potential gradient and higher derivatives, and hence,
3. the very large acceleration of particles and curvature of trajectories,
4. the requirement of very small time-steps of standard integrators (such as the Verlet methods and the Nordsieck–Gear Predictor–Corrector methods^{33,34}).

The KPC algorithm addresses these challenges by first predicting the coordinates and momenta due to the collision with the closest singularity $V^s(r)$ and then correcting the resulting coordinates and momenta according to the residual term, as follows.

Formally, $\mathbf{x}(t)$ and $\mathbf{p}(t)$ are obtained according to the velocity Verlet method:^{33–35}

$$\mathbf{x}(t + \Delta t) = \mathbf{x}(t) + \frac{\Delta t}{m} \mathbf{p}(t) - \frac{\Delta t^2}{2m} \nabla V|_{\mathbf{x}(t)} \quad (2)$$

$$\mathbf{p}(t + \Delta t) = \mathbf{p}(t) - \frac{\Delta t}{2} (\nabla V|_{\mathbf{x}(t)} + \nabla V|_{\mathbf{x}(t+\Delta t)}) \quad (3)$$

where $m = m_e$ is the mass of the particle, and the time increment Δt is assumed to be sufficiently short to ensure energy conservation. However, to address the numerical challenge of the close encounter, we decompose the total force ∇V into the contribution due to the nearest singularity ∇V^s and the contributions due to smaller residual forces ∇V^r and write suggestively:

$$\mathbf{x}(t + \Delta t) = \mathbf{x}(t) + \frac{\Delta t}{m} \mathbf{p}(t) - \frac{\Delta t^2}{2m} (\nabla V^s|_{\mathbf{x}(t)} + \nabla V^r|_{\mathbf{x}(t)}) \quad (4)$$

$$\mathbf{p}(t + \Delta t) = \mathbf{p}(t) - \frac{\Delta t}{2} (\nabla V^r|_{\mathbf{x}(t)} + \nabla V^s|_{\mathbf{x}(t)} + \nabla V^s|_{\mathbf{x}(t+\Delta t)} + \nabla V^r|_{\mathbf{x}(t+\Delta t)}) \quad (5)$$

and introducing auxiliary coordinates

$$\mathbf{x}^r(t) = \mathbf{x}(t) \quad (6)$$

$$\mathbf{p}^r(t) = \mathbf{p}(t) - \frac{\Delta t}{2} \nabla V^r|_{\mathbf{x}(t)} \quad (7)$$

$$\mathbf{x}^s(t + \Delta t) = \mathbf{x}^r(t) + \frac{\Delta t}{m} \mathbf{p}^r(t) - \frac{\Delta t^2}{2m} \nabla V^s|_{\mathbf{x}^r(t)} \quad (8)$$

$$= \mathbf{x}(t) + \frac{\Delta t}{m} \mathbf{p}(t) - \frac{\Delta t^2}{2m} (\nabla V^s|_{\mathbf{x}(t)} + \nabla V^r|_{\mathbf{x}(t)}) \quad (9)$$

$$\mathbf{p}^s(t + \Delta t) = \mathbf{p}^r(t) - \frac{\Delta t}{2} (\nabla V^s|_{\mathbf{x}^r(t)} + \nabla V^s|_{\mathbf{x}^r(t+\Delta t)}) \quad (10)$$

We obtain by comparison with eqs 4 and 5:

$$\mathbf{x}(t + \Delta t) = \mathbf{x}^s(t + \Delta t) \quad (11)$$

$$\mathbf{p}(t + \Delta t) = \mathbf{p}^s(t + \Delta t) - \frac{\Delta t}{2} \nabla V^r|_{\mathbf{x}(t+\Delta t)} \quad (12)$$

The auxiliary variables were constructed such that eqs 8 and 10 become velocity Verlet equations for only the singular potential $V^s(r)$, so they provide a practical way of predicting coordinates and momenta at time $t + \Delta t$, as solely determined by the two-body collision with the closest proton. During a close encounter, however, such equations become numerically stiff, and they are replaced by analytic solutions of the corresponding two-body Kepler initial value problem with initial coordinates $(\mathbf{x}^r(t), \mathbf{p}^r(t))$

[three-body collisions are usually prevented by Coulombic repulsion under cool plasma conditions] as described in section 2.3. [The two-body problem is integrable for a variety of spherically symmetric interaction potentials, including the Coulomb r^{-1} potential and van der Waals r^{-6} type potentials.³² In particular, the two-body problem for the $1/r$ potential is known as the “Kepler problem,” since the elliptical trajectories obey Kepler’s laws of planetary motion.] Having obtained $\mathbf{x}^s(t + \Delta t)$ and $\mathbf{p}^s(t + \Delta t)$, we obtain $\mathbf{x}(t + \Delta t)$ and $\mathbf{p}(t + \Delta t)$ by correcting the predicted coordinates and momenta according to eq 12. It turns out that $\mathbf{x}^r(t)$ and $\mathbf{x}^s(t + \Delta t)$ are not, as indicated in Figure 1, distinct from $\mathbf{x}(t)$ and $\mathbf{x}(t + \Delta t)$, but their corresponding momenta are.

The resulting KPC algorithm thus allows for the integration of close-encounter collisions beyond the capabilities of the standard velocity Verlet algorithm (i.e., eqs 4 and 5), as follows:

A. Determine $\mathbf{p}^r(t)$ by a momentum shift according to eq 6.

B. Obtain $(\mathbf{x}^s(t + \Delta t), \mathbf{p}^s(t + \Delta t))$ by solving the two-body Kepler problem. The analytic solution \mathbf{S}^k is a set that contains position, momentum, and, in principle, all higher derivatives of the position at the final time $t + \Delta t$:

$$(\mathbf{x}^s(t + \Delta t), \mathbf{p}^s(t + \Delta t)) \subset \mathbf{S}^k(\mathbf{x}^r(t), \mathbf{p}^r(t)) \quad (13)$$

C. From $\mathbf{p}^s(t + \Delta t)$, determine $\mathbf{p}(t + \Delta t)$ according to eq 12.

Maximum efficiency is achieved when the analytic task is restricted to small regions surrounding the singularity closest to the scattering particle, since the solution of the Kepler problem is more involved than a velocity Verlet integration step. This is typically ensured by defining a cutoff distance r_{\min} from the singularity center, below which the KPC method is implemented. In the case under consideration, the potential is stationary, and the cutoff distance can be set to a constant value much smaller than half the minimal distance between singularities, $r_{\min} \ll 1/2 \min\{r_{ij}\}$. It should be small enough that the residual potential V^r is always much more slowly varying than the close-encounter potential V^s , guaranteeing that the singular potential dominates the dynamics and the influence of the residual potential leads only to small corrections. To make the method fully time-reversal symmetric, the cutoff criterion has to be made time-reversal symmetric as well, in the following way. The KPC method is used when the initial coordinate is inside the cutoff radius from a scattering center. If the coordinate after the time step is outside the cutoff radius, the result is discarded, and a velocity Verlet step is made instead.

In contrast to ref 31 where a single momentum shift was applied, our symmetrized approach uses two momentum shifts per time step. However, the potential gradient needs only be evaluated once per time step.

2.2. Multibody Molecular Dynamics. The generalization of the KPC algorithm, introduced in section 2.1, to multibody molecular dynamics is straightforward. Let $(\mathbf{x}(t), \xi(t))$ and $(\mathbf{p}(t), \pi(t))$ be the position and momentum vectors of the system at time t . Without a loss of generality, let $(\mathbf{x}(t), \mathbf{p}(t))$ be the six-dimensional phase-space vector describing the relative motion of two particles undergoing a close encounter and m be their reduced mass, while their center-of-mass coordinate, along with the coordinates of all other particles (some of which may also be in a close encounter), is contained in the phase-space vector $(\xi(t), \pi(t))$.

While the interparticle potential $V(\mathbf{x}(t), \xi(t))$ includes all of the interactions, only the dynamics of the relative coordinate

$(\mathbf{x}(t), \mathbf{p}(t))$ are discussed in the following. Accordingly, the gradient will denote the vector of partial derivatives with respect to the relative coordinate \mathbf{x} only:

$$\nabla_3 V = \left(\frac{\partial V}{\partial x_1}, \dots, \frac{\partial V}{\partial x_3} \right) \quad (14)$$

The dependence of the phase-space relative coordinates $(\mathbf{x}(\Delta t), \mathbf{p}(\Delta t))$ on initial conditions $(\mathbf{x}(0), \mathbf{p}(0))$ is then approximated by

$$\mathbf{x}(t + \Delta t) = \mathbf{x}(t) + \frac{\Delta t}{m} \mathbf{p}(t) - \frac{\Delta t^2}{2m} \nabla_3 V|_{(\mathbf{x}(t), \xi(t))} \quad (15)$$

$$\begin{aligned} \mathbf{p}(t + \Delta t) = & \mathbf{p}(t) - \frac{\Delta t}{2} (\nabla_3 V|_{(\mathbf{x}(t), \xi(t))} \\ & + \nabla_3 V|_{(\mathbf{x}(t+\Delta t), \xi(t+\Delta t))}) \end{aligned} \quad (16)$$

To deal with the close encounter, the potential V is again written as a sum of the close encounter potential $V^s(r)$, $r = \|\mathbf{x}\|$, and the residual potential $V^r(\mathbf{x}, \xi)$:

$$V(\mathbf{x}, \xi) = V^s(r) + V^r(\mathbf{x}, \xi) \quad (17)$$

and eqs 15 and 16 become

$$\begin{aligned} \mathbf{x}(t + \Delta t) = & \mathbf{x}(t) \\ & + \frac{\Delta t}{m} \mathbf{p}(t) - \frac{\Delta t^2}{2m} (\nabla_3 V^s|_{\mathbf{x}(t)} \\ & + \nabla_3 V^r|_{(\mathbf{x}(t), \xi(t))}) \end{aligned} \quad (18)$$

$$\begin{aligned} \mathbf{p}(t + \Delta t) = & \mathbf{p}(t) - \frac{\Delta t}{2} (\nabla_3 V^r|_{(\mathbf{x}(t), \xi(t))} \\ & + \nabla_3 V^s|_{\mathbf{x}(t)} + \nabla_3 V^s|_{\mathbf{x}(t+\Delta t)} \\ & + \nabla_3 V^r|_{(\mathbf{x}(t+\Delta t), \xi(t+\Delta t))}) \end{aligned} \quad (19)$$

Following section 2.1, we define appropriate auxiliary coordinates and momenta:

$$\mathbf{x}^r(t + \Delta t) = \mathbf{x}(t) \quad (20)$$

$$\mathbf{p}^r(t + \Delta t) = \mathbf{p}(t) - \frac{\Delta t}{2} \nabla_3 V^r|_{(\mathbf{x}(t), \xi(t))} \quad (21)$$

$$\mathbf{x}^s(t + \Delta t) = \mathbf{x}^r(t) + \frac{\Delta t}{m} \mathbf{p}^r(t) - \frac{\Delta t^2}{2m} \nabla_3 V^s|_{\mathbf{x}^r(t)} \quad (22)$$

$$\mathbf{p}^s(t + \Delta t) = \mathbf{p}^r(t) - \frac{\Delta t}{2} (\nabla_3 V^s|_{\mathbf{x}^r(t)} + \nabla_3 V^s|_{\mathbf{x}^s(t+\Delta t)}) \quad (23)$$

and we obtain

$$\mathbf{x}(t + \Delta t) = \mathbf{x}^s(t + \Delta t) \quad (24)$$

$$\mathbf{p}(t + \Delta t) = \mathbf{p}^s(t + \Delta t) - \frac{\Delta t}{2} \nabla_3 V^r|_{(\mathbf{x}(t+\Delta t), \xi(t+\Delta t))} \quad (25)$$

During a close encounter, eqs 22 and 23 are replaced by the analytic solution of the Kepler problem, as described in section 2.3, and the resulting values are then augmented according to

eqs 24 and 25. The resulting algorithm for multibody molecular dynamics is summarized, as follows:

1. Determine $\mathbf{p}^r(t)$ by momentum shift according to eq 21.
2. Compute $(\mathbf{x}^s(t + \Delta t), \mathbf{p}^s(t + \Delta t))$ by solving the Kepler problem \mathbf{S}^K :

$$(\mathbf{x}^s(t + \Delta t), \mathbf{p}^s(t + \Delta t)) \subset \mathbf{S}^K(\mathbf{x}(t), \mathbf{p}(t)) \quad (26)$$

3. Obtain coordinates $\mathbf{x}(t + \Delta t)$ and $\xi(t + \Delta t)$, as follows:
 - a. From $\mathbf{x}^s(t + \Delta t)$, determine $\mathbf{x}(t + \Delta t)$ according to eq 24.
 - b. Obtain $\xi(t + \Delta t)$ by using velocity Verlet, or otherwise solving steps 2 and 3a for relative coordinates describing two-body close encounters.
4. Obtain momenta $\mathbf{p}(t + \Delta t)$ and $\mathbf{p}^s(t + \Delta t)$, as follows:
 - a. From $\mathbf{p}^s(t + \Delta t)$, determine $\mathbf{p}(t + \Delta t)$ according to eq 25.
 - b. Determine $\pi(t + \Delta t)$, analogously to step 3b.

2.3. Kepler Problem. When a close encounter is detected, we consider the particle attracted by the nearest singularity:

$$V^s(r) = -\frac{\gamma}{r} \quad (27)$$

where

$$\mathbf{r}(t) = \mathbf{x}^s(t) - \mathbf{X} \quad (28)$$

with \mathbf{X} being the position of the singularity. We solve the equation of motion:

$$\ddot{\mathbf{r}} + \mu \frac{\mathbf{r}}{r^3} = 0 \quad (29)$$

with force parameter $\mu = \gamma/m$ resulting from the gravitational or Coulomb coefficient γ and the mass m , and initial conditions $\mathbf{r}_0 = \mathbf{x}(t) - \mathbf{X}$ and $\mathbf{v}_0 = \dot{\mathbf{x}}(t)$. The solutions exploit the conservation of specific angular momentum $\mathbf{L} = \mathbf{r} \times \mathbf{v}$, specific energy $h^s(t) = \mathbf{v}^2/2 - \mu/r$, and eccentricity vector $\mathbf{e} = \mathbf{v} \times (\mathbf{r} \times \mathbf{v})/\mu - \mathbf{r}/r$.

Equation 29 is regularized by introducing the fictitious time τ with

$$\frac{d}{d\tau} = r \frac{d}{dt} \quad (30)$$

leading to the regularized equation of motion

$$\mathbf{r}'' - 2h^s \mathbf{r} = -\mu \mathbf{e} \quad (31)$$

where the derivatives in eq 31 are in respect to τ , and the initial conditions are modified according to eq 30 to $\mathbf{r}(\tau = 0) = \mathbf{r}_0$ and $\mathbf{r}'(\tau = 0) = r\mathbf{v}_0$. This is a harmonic oscillator problem that is readily solved by exponential functions, typically leading to real solutions that are trigonometric or hyperbolic functions.

The motion is characterized according to four classes of possible solutions, including circular, parabolic, elliptic, and hyperbolic, as shown in the following subsections.³⁶ [In contrast, ref 36 considers 13 classes of solutions, including the circular case and 12 other cases generated from the elliptic, hyperbolic, and parabolic classes as subdivided according to the values of the rotational momentum and fictitious time τ , described below.] The circular and parabolic cases have simple, explicit solutions, while the elliptic and hyperbolic cases lead to Kepler equations that need to be solved iteratively.^{37,38}

2.3.1. Circular Motion. When $\mathbf{r}_0 \cdot \mathbf{v}_0 = 0$, the eccentricity $\mathbf{e} = 0$ and eq 31 becomes homogeneous. In this case, $\mathbf{r}_0 \cdot \mathbf{v}_0 = 0$ and $\mathbf{v}_0 = (\mu/r_0)^{1/2}$; i.e., the coordinate change is orthogonal to the

radius vector, so that particle motion is circular and

$$\mathbf{r}(t + \Delta t) = c_r \mathbf{r}_0 + c_v \frac{\mathbf{v}_0}{n} \quad (32)$$

$$\mathbf{v}(t + \Delta t) = -nc_v \mathbf{r}_0 + c_r \mathbf{v}_0 \quad (33)$$

where $c_v = \sin(n(t + \Delta t))$ and $c_r = \cos(n(t + \Delta t))$, with $n = j^3/\mu$ and $j = (2|E^s(t + \Delta t)|/m)^{1/2}$.

2.3.2. *Parabolic Motion.* When $e = 1$, the specific energy $h^s = 0$, and eq 31 has no linear term, giving the parabolic solution

$$\mathbf{r}(t + \Delta t) = \frac{1}{2}(p - \mu[\tau(t + \Delta t)]^2)\mathbf{e} + \tau(t + \Delta t)\mathbf{B} \quad (34)$$

$$r(t + \Delta t) = \frac{1}{2}(p + \mu[\tau(t + \Delta t)]^2) \quad (35)$$

$$\mathbf{v}(t + \Delta t) = \frac{1}{r(t + \Delta t)}(-\mu[\tau(t + \Delta t)]\mathbf{e} + \mathbf{B}) \quad (36)$$

where $\mathbf{B} = \mathbf{L} \times \mathbf{e}$. The *fictitious time* $\tau(t)$, introduced above, is obtained by solving the Kepler equation:

$$t - t_p = \frac{1}{2}\left(p\tau(t) + \frac{\mu}{3}[\tau(t)]^3\right) \quad (37)$$

where t_p is the *pericenter time*:

$$t_p = -\tau_0\left(p + \frac{\mu}{3}\tau_0^2\right) \quad (38)$$

p is the semilatus rectum:

$$p = \frac{L^2}{\mu} \quad (39)$$

and τ_0 is the fictitious time at time t :

$$\tau_0 = \frac{\mathbf{r}_0 \cdot \mathbf{v}_0}{\mu} \quad (40)$$

Equation 37 can be solved explicitly to obtain the fictitious time, as follows:

$$\tau(t) = \frac{1}{\sqrt[3]{\mu}}(\sqrt[3]{t_D + \sqrt{D}} + \sqrt[3]{t_D - \sqrt{D}}) \quad (41)$$

where negative values are assumed for negative arguments of the cube root, $t_D = 3(t - t_p)$ and $D = t_D^2 + p^3/\mu$.

2.3.3. *Elliptic and Hyperbolic Motion.* When $e \neq 0$ and 1, the motion is either elliptic ($e < 1$, $h^s < 0$) or hyperbolic ($e > 1$, $h^s > 0$). In either case, we obtain the eccentric anomaly:

$$\varepsilon(t + \Delta t) = j\tau(t + \Delta t) \quad (42)$$

as the solution of the (elliptic or hyperbolic) Kepler equation at time $t + \Delta t$, as described below. The resulting eccentricity defines the values of c_v and c_r (see below) and, therefore, the coordinates and velocities, as follows:

$$\mathbf{r}(t + \Delta t) = \frac{1}{k}\left(\frac{c_r}{e} - 1\right)\mathbf{e} + \frac{c_v}{ej}\mathbf{B} \quad (43)$$

$$r(t + \Delta t) = \frac{1 - ec_r}{k} \quad (44)$$

$$\mathbf{v}(t + \Delta t) = \frac{1}{er(t + \Delta t)}\left(-\frac{\mu c_v}{j}\mathbf{e} + c_r\mathbf{B}\right) \quad (45)$$

where $k = -2h^s/\mu$ and $j = (2|h^s|)^{1/2}$.

When $h^s(t) < 0$, the eccentricity ε is the solution of the *elliptic Kepler equation*:

$$n(t + \Delta t - t_p) = \varepsilon(t + \Delta t) - e \sin[\varepsilon(t + \Delta t)] \quad (46)$$

$n = j^3/\mu$, which can be solved iteratively, as described in section 2.3.4. The resulting $\varepsilon(t + \Delta t)$ gives the trigonometric functions:

$$c_v = \sin[\varepsilon(t + \Delta t)] \quad (47)$$

$$c_r = \cos[\varepsilon(t + \Delta t)] \quad (48)$$

which determine the coordinates and velocities, according to eqs 43 and 45.

The pericenter time t_p , introduced by eq 46, is obtained from the eccentricity at the initial time ε_0 , as follows:

$$t_p = -\frac{\varepsilon_0 - e \sin \varepsilon_0}{n} \quad (49)$$

where $\varepsilon_0 = \text{atan2}(y, x)$, with

$$y = \sin \varepsilon_0 = \frac{n}{k\mu e}\mathbf{r}_0 \cdot \mathbf{v}_0 \quad (50)$$

$$x = \cos \varepsilon_0 = \frac{1}{e}\left(1 - \frac{nr_0}{j}\right) \quad (51)$$

Analogously, when $h^s(t) > 0$, the eccentricity $\varepsilon(t + \Delta t)$ is the solution of the *hyperbolic Kepler equation*:

$$n(t + \Delta t - t_p) = -\varepsilon(t + \Delta t) + e \sinh \varepsilon(t + \Delta t) \quad (52)$$

which is solved iteratively, as described in section 2.3.4, using the pericenter time:

$$t_p = \frac{1}{n}\left(\sinh^{-1}\left(-\frac{1}{e} \cdot \frac{nr_0 \cdot v_0}{\mu k}\right) + \frac{nr_0 \cdot v_0}{\mu k}\right) \quad (52a)$$

The resulting eccentricity $\varepsilon(t + \Delta t)$ gives the hyperbolic functions

$$c_v = \sinh \varepsilon(t + \Delta t) \quad (53)$$

$$c_r = \cosh \varepsilon(t + \Delta t) \quad (54)$$

that determine the coordinates and velocities, according to eqs 43 and 45.

2.3.4. *Iterative Solution of Elliptic and Hyperbolic Equations.* The elliptic and hyperbolic Kepler equations, introduced by eqs 46 and 52, have the general form

$$n(t + \Delta t - t_p) = M = \pm \varepsilon \mp e \sin(h) \varepsilon \quad (55)$$

When $n(t + \Delta t - t_p) < 0$, M is replaced by its absolute value, as follows:

$$|M| = \pm \varepsilon_a \mp e \sin(h) \varepsilon_a \quad (56)$$

Therefore, the solution of eq 55 is

$$\varepsilon(t + \Delta t) = \varepsilon = \text{sgn}(M)\varepsilon_a \quad (57)$$

Equation 56 is solved iteratively,³⁸ starting from an initial guess $\varepsilon_a^{(0)}$ applicable over the whole range of possible parameters

M and e , where each Halley's iteration is followed by a Newton–Raphson optimization. The iterative scheme typically converges to machine accuracy in about three iterations.³⁷ The initial guess $\varepsilon_a^{(0)}$ is obtained, as follows:

Elliptic Kepler Equation. For elliptic motion ($0 \leq M \leq \pi$ and $0 \leq e \leq 1$) and small M , we expand the *sin* function in eq S6 to third order, and the resulting approximation of M is substituted in eq S5, giving a cubic approximation of the elliptic Kepler equation:

$$0 = \varepsilon_{00}^3 + 3q\varepsilon_{00} - 2r \quad (58)$$

where $q = 2(1 - e)/e$ and $r = 3M/e$. Solving eq 58, we obtain

$$\varepsilon_{00} = \frac{2r}{c_v^2 + q + (q/w)^2} \quad (59)$$

where $c_v = ((r^2 + q^3)^{1/2} + r)^{1/3}$. On the other hand, for large M , a good initial guess is

$$\varepsilon_{01} = M$$

Therefore, we define an initial guess that is valid for intermediate values of M , as an M -weighted average of the small M guess ε_{00} and the large M guess ε_{01} :

$$\varepsilon_a^{(0)} = \frac{1}{\pi}(M \times \varepsilon_{01} + (\pi - M) \times \varepsilon_{00}) \quad (60)$$

$$= \frac{1}{\pi}(M^2 + (\pi - M) \times \varepsilon_{00}) \quad (61)$$

Hyperbolic Kepler Equation. For hyperbolic motion ($M > 0$ and $1 \leq e$) with small M , the *sinh* function in eq S6 is expanded to third order, and the resulting approximation of M is introduced into eq S5 to give the cubic approximation to the hyperbolic Kepler equation:

$$0 = \varepsilon_{00}^3 + 3q\varepsilon_{00} - 2r \quad (62)$$

where $q = 2(e - 1)/e$ and $r = 3M/e$, with the same formal solution introduced by eq S9, although there are different values of q and r .

Bounded coefficients are obtained through an iterative procedure based on the hyperbolic equation divided by e , as follows:

$$0 = -M - \varepsilon + e \sinh[\varepsilon] \quad (63)$$

$$0 = -L - g\varepsilon + \sinh[\varepsilon] \quad (64)$$

with $L = M/e$ and $g = 1/e$. For large L , a good initial guess is

$$\varepsilon_{01} = \sinh^{-1} L \quad (65)$$

which is again mixed with ε_{00} , as follows:

$$\varepsilon_a^{(0)} = \frac{M \times \varepsilon_{01} + 1 \times \varepsilon_{00}}{M + 1} \quad (66)$$

$$= \frac{M \times \sinh^{-1} L + \varepsilon_{00}}{M + 1} \quad (67)$$

3. ELECTRON SCATTERING

3.1. Time-Dependent Wigner Transform. This section illustrates the capabilities of the KPC algorithm, introduced in section 2, as applied to semiclassical dynamics simulations of electron–proton scattering processes in the Wigner-transform

time-dependent picture. Simulations consider the problem of electron scattering from stationary protons, as described within the Born–Oppenheimer approximation (i.e., with $m = m_e$ and the Coulombic parameter $\gamma = 1$ au, so that $\mu = \gamma/m = 1$ au).

The initial state for the scattering electron is defined by the three-dimensional Gaussian:

$$\psi_0(\mathbf{x}) = (2\pi\sigma^2)^{-3/4} \exp\left(-\frac{(\mathbf{x} - \mathbf{x}_0)^2}{4\sigma^2} + \frac{i}{\hbar}\mathbf{p}_0(\mathbf{x} - \mathbf{x}_0)\right) \quad (68)$$

with average position \mathbf{x}_0 and momentum \mathbf{p}_0 . The corresponding Wigner transform:³⁹

$$P_0(\mathbf{x}, \mathbf{p}) = \frac{1}{(2\pi\hbar)^3} \int_{-\infty}^{\infty} \psi_0^*\left(\mathbf{x} + \frac{\mathbf{s}}{2}\right) \psi_0\left(\mathbf{x} - \frac{\mathbf{s}}{2}\right) e^{i\mathbf{p}\cdot\mathbf{s}} d\mathbf{s} \quad (69)$$

$$= \frac{1}{(2\pi\hbar\sigma_x\sigma_p)^3} \exp\left[-\frac{(\mathbf{x} - \mathbf{x}_0)^2}{2\sigma_x^2} - \frac{(\mathbf{p} - \mathbf{p}_0)^2}{2\sigma_p^2}\right] \quad (70)$$

defines the initial phase-space distribution function, where $\sigma_x = \sigma$ and σ_p is defined by the uncertainty relation $\sigma_x\sigma_p = \hbar/2$.

The full quantum-mechanical Wigner distribution $P_t^{\text{QM}}(\mathbf{x}, \mathbf{p})$ is computed as

$$P_t^{\text{QM}}(\mathbf{x}, \mathbf{p}) = \frac{1}{(2\pi\hbar)^3} \int_{-\infty}^{\infty} e^{i/\hbar\mathbf{p}\cdot\mathbf{s}} \psi_t^*\left(\mathbf{x} + \frac{\mathbf{s}}{2}\right) \psi_t\left(\mathbf{x} - \frac{\mathbf{s}}{2}\right) d\mathbf{s} \quad (71)$$

where ψ_t is the solution of the time-dependent Schrödinger equation

$$i\hbar\frac{\partial}{\partial t}\psi_t(\mathbf{x}) = \left(\frac{\hat{\mathbf{p}}^2}{2m_e} + V(\mathbf{x})\right)\psi_t(\mathbf{x}) \quad (72)$$

with

$$V(\mathbf{x}) = -\sum_j \frac{q_j}{|\mathbf{x} - \mathbf{R}_j|} \quad (73)$$

where the sum is over all protons j , with charge $q_j = +e$, and coordinates \mathbf{R}_j . $\psi_t(\mathbf{x})$ is represented on a three-dimensional grid and propagated according to the standard Split Operator Fourier Transform (SOFT) method.^{40,41} The grid is defined as follows: $x_{\alpha k} = x_{\alpha 0} + k\Delta x$, $k = 1, 2, \dots, 128$, where $\alpha = 1-3$ enumerates the Cartesian directions, $x_{10} = -4$ Å, and $x_{20} = x_{30} = -5$ Å. The grid spacings $\Delta x = 10/128$ Å and $\Delta t = 10^{-4}$ fs define a sufficiently fine space–time grid that ensures an accurate representation of the oscillatory structure of $\psi_t(\mathbf{x})$, even during high-energy collisions (e.g., collisions with tens of electronvolts). The full-quantum propagation is based on the short-time Trotter approximation of the time-evolution operator:

$$\psi_{t+\Delta t}(\mathbf{x}) = U(t, t + \Delta t) \psi_t(\mathbf{x}) \quad (74a)$$

$$\approx e^{-i/\hbar V\Delta t/2} e^{-i/\hbar \hat{\mathbf{p}}^2/2m_e\Delta t} e^{-i/\hbar V\Delta t/2} \psi_t(\mathbf{x}) \quad (74)$$

The time-evolved semiclassical Wigner distribution $P_t^{\text{SC}}(\mathbf{x}, \mathbf{p})$ is computed as follows:

$$P_t^{\text{SC}}(\mathbf{x}, \mathbf{p}) = (2\pi)^{-3} \int_{-\infty}^{\infty} d\mathbf{s} \int_{-\infty}^{\infty} d\mathbf{p}_0 \int_{-\infty}^{\infty} d\mathbf{x}_0 e^{i(\mathbf{p} - \mathbf{p}_0)\cdot\mathbf{s}} \delta(\mathbf{x}_t - \mathbf{x}) P_0(\mathbf{x}_0, \mathbf{p}_0) \quad (75a)$$

$$= \frac{1}{N} \sum_{j=1}^N \delta(\mathbf{p} - \mathbf{p}_t(j)) \delta(\mathbf{x}_t(j) - \mathbf{x}) \quad (75)$$

where $\mathbf{x}_t(j)$ and $\mathbf{p}_t(j)$ are coordinates and momenta, obtained by classical KPC propagation. The initial coordinates and momenta

$\mathbf{x}_0(j)$ and $\mathbf{p}_0(j)$ are sampled by Box–Muller Monte Carlo,⁴² using the phase-space distribution $|P_0(\mathbf{x}_0, \mathbf{p}_0)|$.

3.2. Results. Three model systems were analyzed, including electron scattering from a single central proton (model I), scattering from a central proton in the presence of a peripheral proton (model II), and scattering through a cluster of 125 protons in a configuration typical of a high-density plasma (model III). In models I and II, the initial state for the scattering electron was defined according to eq 68, with $\sigma_0 = \sigma = 0.5$ Å, $x_0 = -1$ au, $z_0 = 0$, and $y_0 = k \times 0.2$ au, where $k = 1, 2, \dots, 5$. Therefore, the initial momentum of the scattering electron was defined as follows:

$$\begin{aligned} p_0^2 &= 2m_e \langle T_0 \rangle \\ &= 2m_e (E - \langle V_0 \rangle) \\ &\approx 2m_e \left(E + \frac{1}{r_0} \right) \end{aligned} \quad (76)$$

with $r_0^2 = x_0^2 + y_0^2$, and $E = 9.2$ eV, defining the initial kinetic energies as listed in Table 1.

Figure 2 shows the comparison of electron–proton scattering trajectories, as described by a single classical trajectory (dots, with initial coordinates and momenta defined by the expectation values of the initial state) and the corresponding full-quantum (SOFT, crosses) and Wigner semiclassical expectation values (lines).

Table 1. Impact Parameters y_0 and Initial Kinetic Energies (K.E.) for Trajectories Shown in Figure 2

y_0 (a.u.)	initial K.E. (eV)
0.2	35.9
0.4	34.5
0.6	32.5
0.8	30.4
1.0	28.4

Figure 2 shows that classical trajectories and benchmark full-quantum trajectories agree at very early times but quickly deviate from each other. In contrast, the semiclassical Wigner description is in almost quantitative agreement with full quantum dynamics throughout the whole propagation time for all cases investigated, including model II where scattering trajectories curve away from the central proton due to the significant influence of the peripheral scattering center and the nearly symmetric impact of the central proton on the extended wave packet.

The origin of small deviations, shown in Figure 2, when comparing the Wigner semiclassical description to the full-quantum results, can be traced to the comparison of the distribution functions in configurational space (see Figure 3). The initial (left) and final (right) densities are shown for impact parameter 0.4 Å for models I (top) and II (bottom), respectively. Large dots indicate proton positions, while small dots correspond to the ensemble distribution. Contours are drawn at σ , 2σ , and 3σ from the maximum density integrated over the z coordinate (60.6%, 13.5%, and 1.1%). SOFT (red) density distributions are compared to the semiclassical Wigner distributions (blue), collected in 64^2 quadratic bins covering the quantum grid (i.e., each bin covers 2^3 quantum grid cells). For illustration purposes, the ensemble of trajectories shown in Figure 2 corresponds to a simulation using $5^6 = 15\,625$ trajectories. Contour lines and quantitative measures are derived from simulations using $7^6 = 117\,649$ trajectories. A maximum allowed energy change of 2.72×10^{-2} eV/fs for each trajectory was enforced at each time step as the basis for the adaptive time step, with a smallest allowed time step of 10^{-39} s.

Deviations between SOFT and semiclassical results, shown in Figure 3, include small components of the semiclassical distributions that remain bound, localized at the protons. This is observed even at the final propagation time, although the full quantum distributions have no bound components. This is an intrinsic limitation of the semiclassical Wigner transform picture that becomes even more pronounced for lower energy collisions, when there are more initial conditions bound in the Coulombic

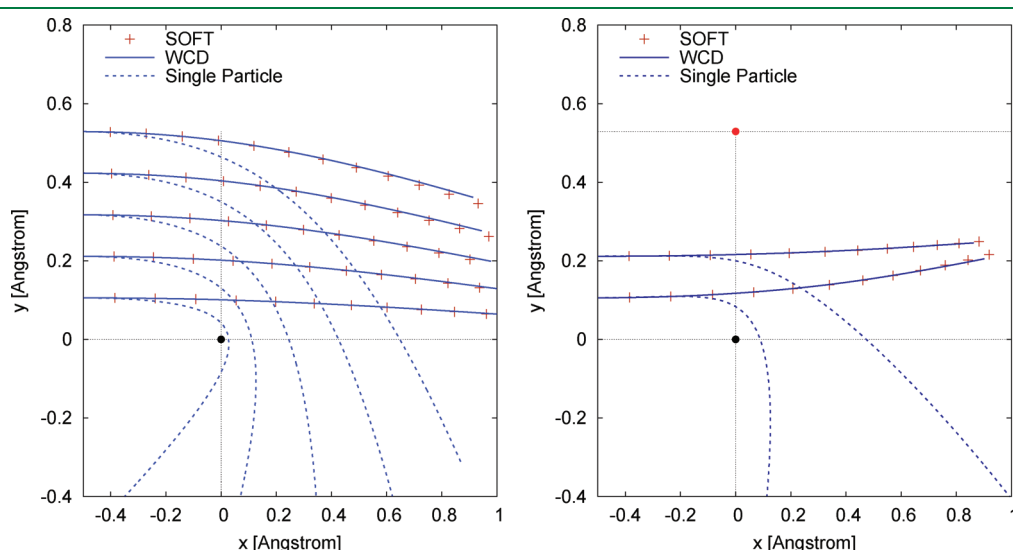


Figure 2. Electron scattering trajectories obtained by expectation values of SOFT full-quantum (red crosses) propagation Wigner classical dynamics (WCD, solid blue) and classical propagation of a single trajectory with initial position and momentum as defined by the expectation values of the initial wave packet (blue dots). Left panel (model I): electron collision with a single proton (black bullet) at the origin. Right panel (model II): collision with two protons, including a central proton at the origin (black bullet) and a peripheral proton (red bullet) at (0,1) au.

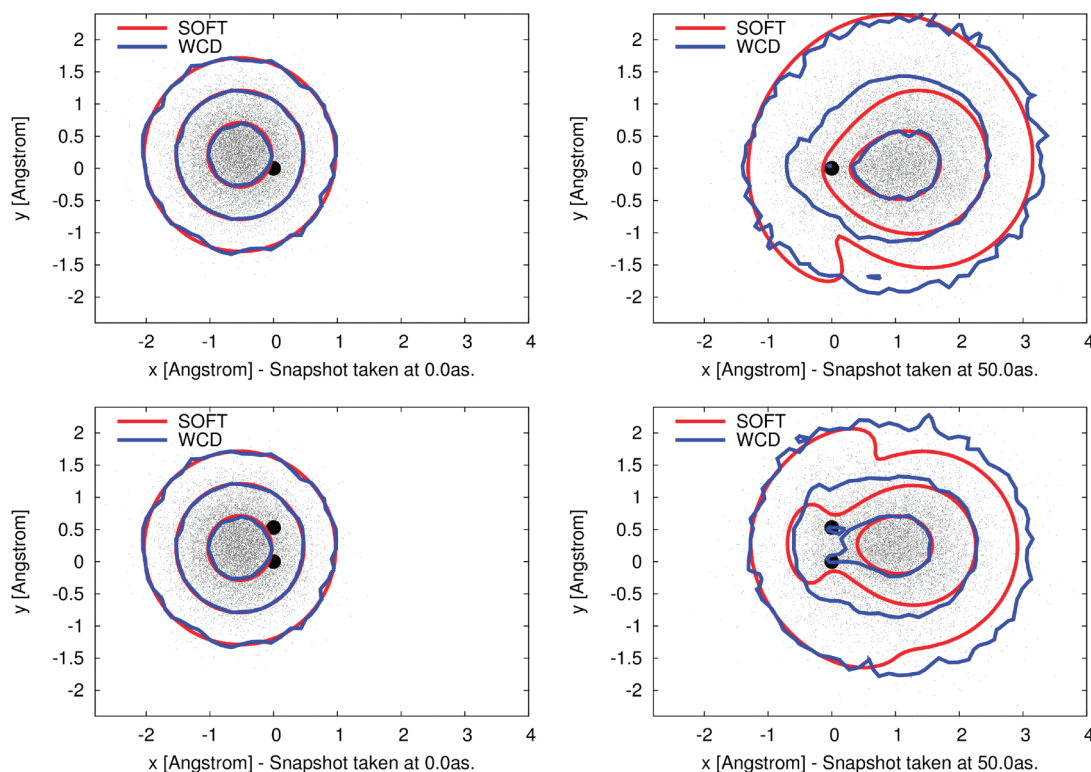


Figure 3. Initial (left) and final (right) densities for impact parameter 0.4 \AA for electron scattering in models I (top) and II (bottom), respectively. Large dots indicate proton positions; small dots are WCD representative configurations. Contours are drawn at percentages of the maximum density (integrated over the z coordinate) found at σ , $2 \times \sigma$, and $3 \times \sigma$ from the center of the distribution, i.e., 60.65%, 13.5%, and 1.1%. Color key: SOFT (red), WCD (blue).

well. As a result, the position predicted by the Wigner transform lags behind the quantum result, faintly visible in Figure 2.

The convex features of the final quantum densities are reproduced well by the WCD method, while concave features in the lowest contour level of the quantum density are due to interference effects and by construction not present in WCD. Nevertheless, the semiclassical Wigner transform reproduces the overall features of the quantum distribution. In fact, a quantitative analysis of the normalized distributions shows $>92\%$ overlap between the semiclassical and quantum distributions for all cases investigated. Even the time-dependent widths, describing the anisotropy of the distribution functions, are in good qualitative agreement with full quantum results.

Figure 4 shows the widths for each of the Cartesian directions describing the time-dependent anisotropy of the distributions. Note that both quantum and semiclassical results show more delocalization along the x direction than in the orthogonal directions y and z . This is likely due to the head-on collision causing the wave packet to undergo more significant deformation in the direction of propagation.

Figure 4 shows that the semiclassical distributions slightly overestimate the widths since they miss interference effects leading to partial localization of the quantum wave packet. This is most prominent in the x direction due to the bound component of the semiclassical distributions, although the trends and overall agreement are quite satisfactory. In fact, close inspection of Figure 3 shows that the semiclassical dynamics reproduce the full-quantum distributions very well, while featuring bound components and deviations at the lowest-density contour level. The concave features in the final quantum density are due to

interference effects, which by construction are not present in WCD. At electron energies above 1 keV, interference becomes negligible, and the agreement of WCD with quantum results becomes excellent. Deviations in the long tails of the distributions, however, affect the overall widths σ disproportionately.

Analogous results are obtained for the description of electron scattering through a cluster of protons (model III). Figure 5 shows the semiclassical (blue) and quantum (red) distributions for a high-energy collision of an electron passing through a disordered cluster of 125 protons (black dots, shown larger for protons closer to the $z = 0$ plane), at the initial (left, $t = 0$ as) and final (right, $t = 50$ as) propagation times. The configuration of the cluster,⁴³ contained in a box with dimensions $5 \times 5 \times 5 \text{ \AA}$, has been extracted from a plasma of density $\rho = 10^{24} \text{ cm}^{-3}$. The initial state for the scattering electron is defined with a width according to a 1s state of a hydrogen atom, and with initial kinetic energy $p_0^2/2m_e = 250 \text{ eV}$.

Numerical Effort. Wall times for production run calculations on a 2.67 GHz intel Core i7 CPU are shown in Table 2. KPC calculations (second column) are compared to results obtained according to the adaptive velocity Verlet method (third column) for two sets of trajectories. The total simulation time is 50 as ($5 \times 10^{-17} \text{ s}$). A maximum of 2^{20} subdivisions of the default time step was allowed, after which a trajectory was marked as failed if it did not satisfy a maximum allowed energy change of $2.72 \times 10^{-2} \text{ eV/fs}$ in one default time step. Trajectories do not fail for the KPC method, while a complete treatment of the failed trajectories in the velocity Verlet method require longer times than given here or are impossible altogether. A larger number of trajectories with randomized initial conditions means an increase

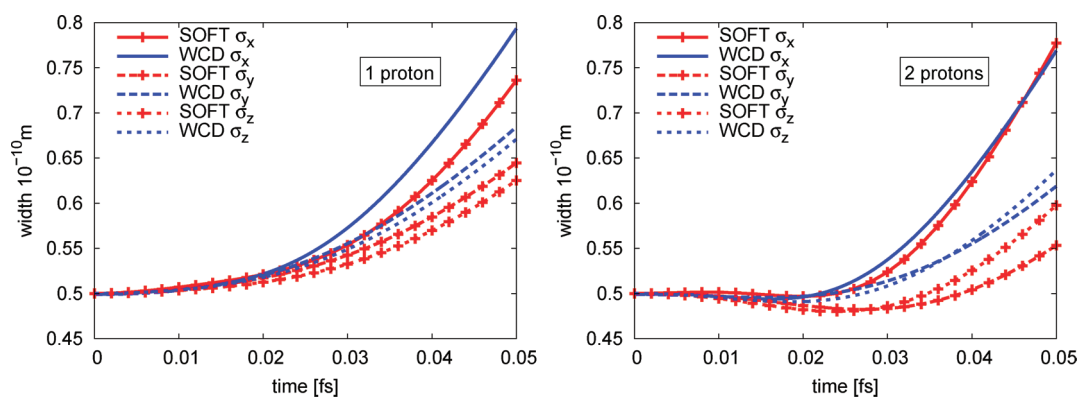


Figure 4. Time dependence of the widths of the time-dependent distributions, as described by semiclassical and quantum calculations of electron scattering in model I (left) and model II (right), respectively. Semiclassical Wigner transforms tend to overestimate the widths, since they lack interference terms responsible for partial localization of the full quantum distributions.

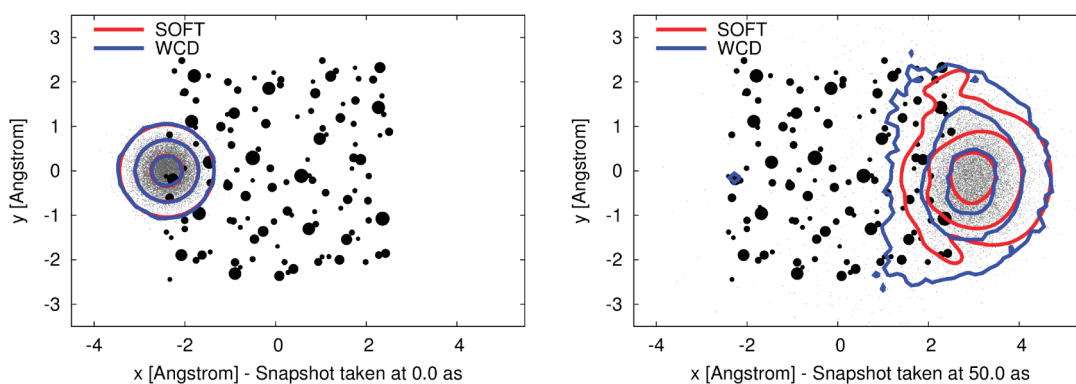


Figure 5. Contour plots of quantum (red) and semiclassical (blue) probability densities, integrated over the z coordinate at the initial (left, $t = 0$ as) and final (right, $t = 50$ as) propagation times, for a high-energy collision of an electron passing through a disordered cluster of 125 protons (black dots, shown larger for protons closer to the $z = 0$ plane). Coordinates x and y in Å.

Table 2. Wall Times for Production Run Calculations (in hours:minutes:seconds) on a 2.67 GHz Intel Core i7 CPU of Adaptive KPC Method with Cutoff (Second Column) and Adaptive Velocity Verlet Method (Third Column) for Two Different Numbers of Trajectories for a Total Simulation Time of $50 \text{ as} \times 10^{-17} \text{ s}$

particles	adaptive		failed KPC trajectories	failed Verlet trajectories
	KPC	adaptive Verlet		
15625	17:25	23:00 (+32%)	0	1 (0.0064%)
117649	1:55:04	2:44:33 (+43%)	0	15 (0.013%)

in the probability of close encounters, which is reflected by the increased ratio between the run times of the two methods and the increased percentage of failed particles.

The selection of trajectories that fail to conserve energy to within $2.72 \times 10^{-2} \text{ eV/fs}$ when propagated according to velocity Verlet are uncontrolled by the user and depend on the dynamics, so that the statistics of the result are skewed. The number of close encounters scales with the third root of the number of particles and linearly with time, making failing trajectories a considerable problem for larger simulations. The adaptive KPC method with cutoffs is more efficient and suffers from no such drawback, so that production runs with 1 million or more trajectories can be performed routinely.

Numerical Accuracy. All KPC results discussed above were performed with an adaptive time step, ensuring that energy conservation is satisfied to a given accuracy. The absence of failed trajectories for the KPC method shows that, time step for time step, the KPC method is more accurate than velocity Verlet and effectively solves the close encounter problem.

To quantify energy conservation, close encounter simulations at a constant time step of a single particle with a resting proton were performed, with a peripheral proton at 1 au distance (model II). Figure 6 shows the energy change of KPC at a constant time step for hyperbolic (left) and elliptic (right) character trajectories. KPC results are shown for cutoff radii of 0.3 au (black lines), 0.2 au (dark gray lines), 0.1 au (light gray lines), and no cutoff, i.e., never using velocity Verlet at any distance (dashed lines). The difficulty of the close encounter is expressed by energy nonconservation caused by it. As the electronic particle approaches the protonic scattering center, the velocity Verlet energy deviates from its initial value, and when switching to the KPC method at the cutoff radius, energy is conserved again before oscillating at the point of closest approach. The energy then returns to the value before its oscillation, before traversing the cutoff radius causes a switch back to velocity Verlet. This demonstrates that, at a given time step, the KPC method conserves energy much closer to the scattering center than velocity Verlet.

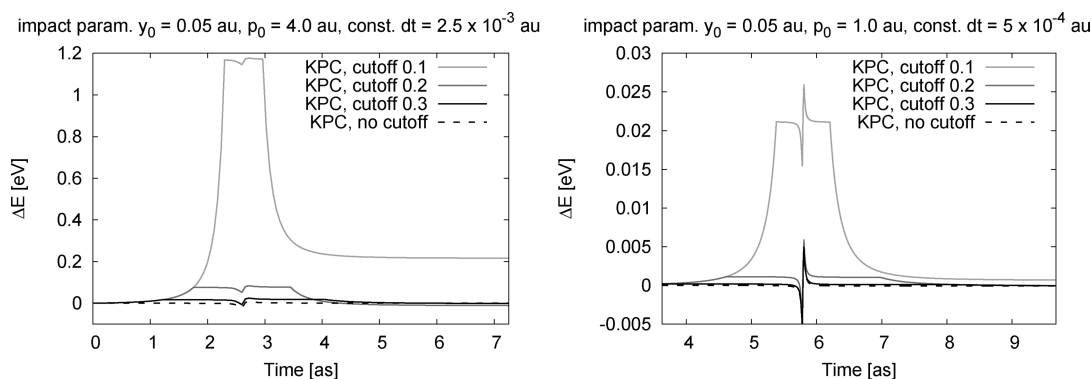


Figure 6. Energy change of KPC at constant time step for hyperbolic (left) and elliptic (right) character trajectories. KPC results are shown for cutoff radii of 0.3 au (black lines), 0.2 au (dark gray lines), 0.1 au (light gray lines), and no cutoff, i.e. never using velocity Verlet at any distance (dashed lines). It can be seen that, approaching the scattering center, the velocity Verlet shows a significant energy deviation, while switching to KPC causes the energy to level off. Energy nonconservation at the smaller cutoff of 0.1 is due to a break down of the velocity Verlet method already outside the cutoff radius, while using KPC only leads to well conserved energy.

At large cutoff radii, the energy eventually returns to its initial value, demonstrating that the time-reversal symmetric construction of the KPC method solves the close-encounter problem. At the smaller cutoff radius of 0.1 au, the energies do not quite return to their initial value which, considering the success at larger cutoffs, must be ascribed to a break down of the velocity Verlet method already outside the cutoff radius. Velocity Verlet by itself fails at the given time steps, as witnessed by macroscopic energy shifts after the close encounters of 0.30 au (8.3 eV) and 0.55 au (14.9 eV). Note that a shorter time step is shown for the elliptic case, as that leads to a smaller distance of closest approach. For smaller time steps, accuracy is improved, while for larger time steps, the shortcomings of velocity Verlet become more pronounced.

It is worth mentioning that when adaptive time steps are used, smaller cutoff values may be more useful since they allow for a reduction in the numerical effort, which per time step is larger for the KPC method, as well as making sure that cutoff spheres between neighboring protons never overlap, which is why 0.1 au has been used for all adaptive time step Wigner density propagations above. In calculations with many potential wells, where the integrator effort is negligible in comparison to that for evaluation of the potential gradient, a larger cutoff is more useful.

4. DISCUSSION

The KPC method is a highly efficient multipurpose method for simulations of an important class of dynamical problems featuring singular potentials. The method is easy to implement and offers an accurate treatment of dynamical problems in which close encounters between mutually attractive particles may occur. For comparable integration time steps, the KPC method is more accurate than standard integrators, even when the Kepler equation is solved to machine accuracy, since its accuracy is determined by the impact of the slowly varying residual potential on the trajectory (and its occasionally sudden changes of direction). A sample program reproducing the calculations reported in Figure 6 is available free of charge upon request to the corresponding authors.

The KPC method can be made arbitrarily accurate, as opposed to standard integrators which typically break at some finite distance of closest approach when the required time step becomes so small that the operations involved cannot be computed at the given machine accuracy. The cost of the KPC method increases when the distance between singularities decreases,

since the cutoff distance must be reduced to keep it smaller than half the minimal distance between singularities. As the cutoff is reduced, the cost of standard propagation methods, applied outside the cutoff distance, increases.

As presented in this paper, the KPC method is limited to problems with close encounters of two-body collisions. This is usually sufficient for most molecular dynamics simulations where the Coulomb repulsion limits close encounters to pairs of particles of opposite charge and prevents three- and higher-body collisions. The method, however, could still be used in applications to gravitational dynamics where multibody close encounters are much less common than two-body collisions.

To make the KPC method time-reversal symmetric, it has to be made sure that the cutoff distance is crossed during the same default time step in both directions. This is done by observing whether the cutoff is crossed from the inside to the outside during the default time step, reverting back to the underlying numerical integrator if this occurs.

As illustrated for models I–III, the KPC method allows for semiclassical dynamics simulations of phase-space distributions in very good agreement with quantum dynamics simulations. For electron scattering from attractive Coulomb potentials, the KPC approach provides a highly parallelizable approach. The method can be applied in conjunction with a wide range of standard integrators, including high-order predictor corrector methods (such as the Nordsieck–Gear method) since the solution S^K of the Kepler problem yields also higher time derivatives of the position that can be correctly augmented, as described by eq 25.

At large times, the sampling of classical trajectories becomes an issue, as the dispersion predicted by the Wigner trajectories is limited by the largest momentum among these trajectories. The predictive power of the Wigner density is also easily seen to be limited in its spatial resolution by the number of trajectories employed. In the multiple scattering case, self-interference of the electron may become significant after many interactions, which is not represented by the current method. In the cases presented here, it turns out to play a minor role not significantly altering the resulting densities, so that the Wigner trajectory method yields good agreement with SOFT.

As discussed in previous sections, the KPC method has been implemented and illustrated as applied to modeling single electron scattering from unscreened Coulombic potentials, using

Table 3. Percentage Deviations of the Debye–Hückel Potential and Its Gradient, Relative to the Unscreened Coulomb Potential, for Different Values of Debye Screening Distances and Cutoff Radii

relative deviation	$r = r_{\min}^{\text{Verlet}} = 10^{-2}$ au			$r = r_{\min}^{\text{Gear}} = 10^{-3}$ au		
	$d = 5$ au	$d = 10$ au	$d = 20$ au	$d = 5$ au	$d = 10$ au	$d = 20$ au
$\delta V^{\text{DH}}(r)$	2×10^{-3}	10^{-3}	5×10^{-4}	2×10^{-4}	10^{-4}	5×10^{-5}
$\delta \nabla V^{\text{DH}}(r)$	2×10^{-6}	5×10^{-7}	1.25×10^{-7}	2×10^{-8}	5×10^{-9}	1.25×10^{-9}

a cutoff radius r_{\min} around the scattering centers chosen to optimize accuracy and performance. The multielectron scattering problem becomes rather complicated at low energies due to electron exchange and correlation. Therefore, the Kepler predictor corrector method is expected to be most useful for the description of electron scattering processes at high (keV) energies. In this case, fast electrons may be seen as dressing the nuclear potential seen by an additional electron in the form of the Yukawa, or Debye–Hückel, screened-Coulomb potential:⁴⁴

$$V^{\text{DH}}(r) = \frac{1}{r} \exp\left(-\frac{r}{d}\right) \quad (77)$$

where d is the Debye screening distance. Analogous treatments of KPC can be applied for this and other classes of spherical potentials, especially those whose two-body initial value problem is solved analytically.

The Yukawa potential in particular, however, converges to the Coulomb potential when $r \rightarrow 0$. Therefore, for a sufficiently small r_{\min} , the solution of the unscreened Kepler problem already gives a good approximation to close encounters for the Yukawa potential. The intrinsic error can be minimized by setting r_{\min} to an appropriate value. Table 3 shows how small are the percentage deviations of the Debye–Hückel potential and its gradient, when compared to the unscreened Coulomb potential, for various screening distances d and cutoff radii r_{\min} .

The deviations can be approximated by Taylor expansion of the exponential:

$$\delta V^{\text{DH}}(r) = \frac{V^{\text{DH}}(r) - V^{\text{s}}(r)}{V^{\text{s}}(r)} \approx \frac{r}{d} \quad (78)$$

$$\delta \nabla V^{\text{DH}}(r) = \frac{|\nabla V^{\text{DH}}(r) - \nabla V^{\text{s}}(r)|}{|\nabla V^{\text{s}}(r)|} \approx \frac{1}{2} \left(\frac{r}{d}\right)^2 \quad (79)$$

Through an appropriate choice of r_{\min} , any desired accuracy may be achieved for the KPC method for any given d , with a corresponding impact on computational performance if small r_{\min} values are chosen. Careful analysis of the numerical effort shows that with the Verlet method, a cutoff of 10^{-2} au may be chosen with only minimal penalty to the efficiency of the KPC method. Under ignition conditions, proton density is on the order of 10^{26} cm^{-3} , where the mean distance between protons is larger than 0.2 au, i.e., much larger than this cutoff. If better numerical accuracy is required, efficiency may be traded in to obtain even smaller cutoff distances.

Another class of potentials is the repulsive Coulombic potential between protons, commonly found in high energy density plasma simulations where kinetic energies are sufficiently high as to cause close encounters between protons. As at kiloelectronvolt energies, electron exchange and correlation play a subordinate role for the system's dynamics, the single electron molecular dynamics methods proposed should be applicable. For such simulations, a

repulsive KPC method can be constructed analogously from the analytic solution of the repulsive Kepler problem. For lower energies (the warm dense matter regime), alternative methods taking exchange and correlation into account have to be explored.

5. CONCLUSIONS

We have introduced the KPC algorithm for accurate and efficient simulations of dynamics of particles with attractive $1/r$ singular potentials. When used in its time-reversible form (with a carefully chosen cutoff radius around singularities), the KPC method always reduces the numerical effort with respect to standard integrators and allows for the description of close encounter collisions. The method is easy to implement and should be practical for a wide range of applications where particles gravitate into each other, such as electron–proton interactions and ionic dynamics, as well as applications in other fields with similar computational challenges such as molecular dynamics of high-density plasmas and celestial mechanics.

We have shown how to apply the KPC method to model semiclassical dynamics of electron–proton scattering processes in the Wigner-transform time-dependent picture. The reported results show excellent agreement with benchmark quantum dynamics calculations, including models with multiple scattering centers that defy the capabilities of standard integration methods. The reported results suggest that the Wigner semiclassical dynamics is a practical and accurate approach to include quantum effects in high energy electron–proton collisions when simulated according to the KPC method. The KPC method's applicability to other singular potentials featuring close encounters should provide a useful, easy to implement tool for a wide range of studies, including electron–ion scattering events and particle–antiparticle dynamics, as well as in classical simulations of charged interstellar gas dynamics and gravitational celestial mechanics, where the latter has not been able to profit from KS-regularization.

AUTHOR INFORMATION

Corresponding Author

*E-mail: andreas.markmann@yale.edu; victor.batista@yale.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

V.S.B. acknowledges supercomputer time from NERSC and support from Lawrence Livermore National Laboratory, grant B590847. The NSF grants CHE-0911520 and ECCS-0404191 supported the development of methods for quantum dynamics simulations. We thank Michael Surh, David Richardson, and Jim Glosli at Lawrence Livermore National Laboratory and Paul Grabowski and Michael Murillo at Los Alamos National Laboratory for valuable comments.

■ REFERENCES

- (1) Jortner, J.; Bixon, M. *Electron Transfer - from Isolated Molecules to Biomolecules*; John Wiley and Sons, Inc.: New York, 1999; pp 1.
- (2) Schwartz, B.; Rossky, P. *J. Chem. Phys.* **1994**, *101*, 6902–6916.
- (3) Turi, L.; Sheu, W.; Rossky, P. *Science* **2005**, *309*, 914–917.
- (4) Harumiya, K.; Kawata, I.; Kono, H.; Fujimura, Y. *J. Chem. Phys.* **2000**, *113*, 8953–8960.
- (5) Dunn, T.; Broyles, A. *Phys. Rev.* **1967**, *157*, 156–166.
- (6) Ashcroft, N. J. *Phys. C: Proc. Phys. Soc.* **1968**, 232–243.
- (7) Turi, L.; Borgis, D. *J. Chem. Phys.* **2002**, *117*, 6186–6195.
- (8) Singh, S.; Kumar, S.; Srivastava, M. *J. Phys. B: Atom. Mol. Phys.* **1978**, *11*, 3061–3066.
- (9) Filinov, A. V.; Golubnychiy, V.; Bonitz, M.; Ebeling, W.; Dufty, J. *Phys. Rev. E* **2004**, *70*, 046411.
- (10) Kimura, M.; Inokuti, M. *Phys. Rev. A* **1988**, *38*, 3801–3803.
- (11) Avinash, K.; Eliasson, B.; Shukla, P. *Phys. Lett. A* **2006**, *353*, 105–108.
- (12) Heggie, D.; Hut, P. *The Gravitational Million-Body Problem*; Cambridge University Press: Cambridge, U.K., 2003; p 1.
- (13) Chambers, J. E. *Mon. Not. R. Astron. Soc.* **1999**, *304*, 793–799.
- (14) Heggie, D. *Introduction to stellar dynamics and N-body integrators*. http://manybody.org/modest/heggie_split.pdf (accessed Nov. 2011).
- (15) Paolini, F.; Cabral, E.; dos Santos, A. *36th EPS Conf. Plasma Phys. Sofia* **2009**, *33E*, O–4.039.
- (16) Hernquist, L.; Hut, P.; Makino, J. *Astrophys. J.* **1993**, *402*, L85–L88.
- (17) Kustaanheimo, P.; Stiefel, E. *J. Reine Angew. Math.* **1965**, *218*, 204–219.
- (18) Vivarelli, M. D. *Celest. Mech. Dyn. Astron.* **1985**, *36*, 349–364.
- (19) Neusch, W. *Quaternionic regularisation of perturbed Kepler motion*; Preprint, 1991.
- (20) Aarseth, S. Direct methods for N-body simulations. In *Multiple Time Scales*; Brackbill, J., Cohen, B., Eds.; Academic Press: New York, 1985; pp 377–418.
- (21) Aarseth, S. J. *Gravitational N-Body Simulations*; Cambridge University Press: Cambridge, U.K., 2003; Cambridge monographs on mathematical physics, p 1.
- (22) Aarseth, S. *N-Body Simulation Software*. <http://www.ast.cam.ac.uk/~sverre/> (accessed Nov. 2011).
- (23) Kleppner, D.; Kolenkow, R. J. *An Introduction to Mechanics*, 1st ed.; Cambridge University Press: Cambridge, U.K., 2010; pp 414–416.
- (24) Tuckerman, M.; Martyna, G.; Berne, B. *J. Chem. Phys.* **1990**, *93*, 1287.
- (25) Tuckerman, M.; Berne, B.; Martyna, G. *J. Chem. Phys.* **1992**, *97*, 1990.
- (26) Janezic, D.; Merzel, F. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 321–32.
- (27) Janezic, D.; Merzel, F. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1048–1054.
- (28) Janezic, D.; Praprotnik, M. *Int. J. Quantum Chem.* **2001**, *84*, 2–12.
- (29) Janezic, D.; Praprotnik, M. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1922–1927.
- (30) Miyamoto, S.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 952–962.
- (31) Zwignagel, G.; Toepffer, C.; Reinhard, P.-G. *Phys. Rep.* **1999**, *309*, 117–208.
- (32) Broucke, R. *Astrophys. Space Sci.* **1980**, *72*, 33–53.
- (33) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford Science Publications: Oxford, 1987; p 25.
- (34) Vesely, F. J. *Computational Physics - An Introduction*, 2nd ed.; Kluwer Academic/Plenum Publishers: New York–London, 2001; p 107.
- (35) Swope, W. C.; Andersen, H.; Berens, P. H.; Wilson, K. R. *J. Chem. Phys.* **1982**, *76*, 637–649.
- (36) Condurache, D.; Martinusi, V. *Meccanica* **2007**, *42*, 465–476.
- (37) Odell, A. W.; Gooding, R. H. *Celestial Mechanics* **1986**, *38*, 307–334.
- (38) Gooding, R. H.; Odell, A. W. *Celestial Mechanics* **1988**, *44*, 267–282.
- (39) Wigner, E. *Phys. Rev.* **1932**, *40*, 949–759.
- (40) Feit, M. D.; J. A. Fleck, J.; Steiger, A. *J. Comput. Phys.* **1982**, *47*, 412–433.
- (41) Feit, M. D.; J. A. Fleck, J. *J. Chem. Phys.* **1983**, *78*, 301.
- (42) Box, G.; Muller, M. E. *Ann. Math. Stat.* **1958**, *29*, 610–611.
- (43) Graziani, F. R.; Batista, V. S.; Benedict, L. X.; Castor, J. I.; Chen, H.; Chen, S. N.; Fichtl, C. A.; Glosli, J. N.; Grabowski, P. E.; Graf, A. T.; Hau-Riege, S. P.; Hazi, A. U.; Khairallah, S. A.; Krauss, L.; Langdon, A. B.; London, R. A.; Markmann, A.; Murillo, M. S.; Richards, D. F.; Scott, H. A.; Shepherd, R.; Stanton, L. G.; Streitz, F. H.; Surh, M. P.; Weisheit, J. C.; Whitley, H. D. *High Energy Dens. Phys.* **2011**; DOI: 10.1016/j.hedp.2011.06.010.
- (44) Iafate, G. J.; Mendelsohn, L. B. *Phys. Rev.* **1969**, *182*, 244–258.

Constant pH Molecular Dynamics Simulations of Nucleic Acids in Explicit Solvent

Garrett B. Goh,[†] Jennifer L. Knight,[†] and Charles L. Brooks^{*,†,‡}

[†]Department of Chemistry and [‡]Biophysics Program, University of Michigan, 930 N. University, Ann Arbor, Michigan 48109, United States

S Supporting Information

ABSTRACT: The nucleosides of adenine and cytosine have pK_a values of 3.50 and 4.08, respectively, and are assumed to be unprotonated under physiological conditions. However, evidence from recent NMR and X-ray crystallography studies has revealed the prevalence of protonated adenine and cytosine in RNA macromolecules. Such nucleotides with elevated pK_a values may play a role in stabilizing RNA structure and participate in the mechanism of ribozyme catalysis. With the work presented here, we establish the framework and demonstrate the first constant pH MD simulations (CPHMD) for nucleic acids in explicit solvent, in which the protonation state is coupled to the dynamical evolution of the RNA system via λ -dynamics. We adopt the new functional form $\lambda^{N_{\text{exp}}}$ for λ that was recently developed for multisite λ -dynamics (MS λ D) and demonstrate good sampling characteristics in which rapid and frequent transitions between the protonated and unprotonated states at $\text{pH} = pK_a$ are achieved. Our calculated pK_a values of simple nucleotides are in a good agreement with experimentally measured values, with a mean absolute error of 0.24 pK_a units. This work demonstrates that CPHMD can be used as a powerful tool to investigate pH-dependent biological properties of RNA macromolecules.

1. INTRODUCTION

An increasing number of experimental studies in recent years have recognized the role of protonated nucleotides, particularly adenine and cytosine, in RNA structure and function. Experimental pK_a values of the nucleosides have been measured to be 3.50 for adenosine and 4.08 for cytidine, which become protonated at the N1 and N3 atoms, respectively (Figure 1).¹ These findings suggest that adenine and cytosine should typically be unprotonated at physiological pH and that their contributions to RNA structure and function were initially assumed to be minimal. However, recent studies have revealed the prevalence of protonated adenine and cytosine in a wide variety of nucleic acid structures ranging from DNA triple helices to the anticodon stem loop of tRNA,^{2–6} indicating that the pK_a value of these residues may be shifted upward to near physiological conditions of pH 7. Protonated bases have been reported to be responsible for a number of noncanonical base pair configurations, which suggests their ability to influence RNA structure.^{7,8} A key example is the wobble $A^+ \cdot C$ base pair that has been implicated in stabilizing RNA loop structures^{9,10} and in the pH-dependent conformational flexibility of the ribosomal peptidyl transferase center.¹¹ Wobble $A^+ \cdot C$ base pairs may also form in DNA under biologically relevant conditions,¹² where they can have mutagenic and carcinogenic effects.^{13,14} Apart from these structural influences, it has been suggested that the elevated pK_a of these nucleic acids may play a significant role in ribozyme catalysis, as these protonated residues may be involved in general acid–base catalysis, playing an analogous role to histidine residues in proteins.^{15–19} Some examples of its role in catalysis include hepatitis δ virus ribozyme^{20–24} and hairpin ribozyme,^{25–30} where experimental studies have demonstrated that a loss-of-function mutation of key residues that have elevated pK_a values led to a significant drop in catalytic activity.

Despite the copious amount of biochemical and structural studies that are available for these RNA structures, there still remains some ambiguity as to the exact function of these protonated residues. For example, while experimental studies strongly suggest that adenine 38 (A_{38}) participates in the cleavage and ligation reaction that is catalyzed by the hairpin ribozyme, its specific role in the catalytic mechanism and the structural dynamics of the ribozyme remains disputed.^{25–30} In such situations, *in silico* modeling of RNA structures may shed some light on the existing controversy. Walter and co-workers have demonstrated the usefulness of using molecular dynamics (MD) simulations to clarify the role of the protonated A_{38} in the hairpin ribozyme by suggesting that it serves as a general acid in aligning reactive groups and stabilizing the negative charge.^{31,32} However, such traditional MD simulations are limited in the sense that prior knowledge obtained from experiment about the identity of key catalytic residue(s) and its protonation state(s) is required. In terms of *in silico* prediction of pK_a values, Honig and co-workers have recently demonstrated the ability to accurately calculate the pK_a values of nucleotides using numerical solutions to the Poisson–Boltzmann equation from a series of representative static snapshots obtained from RNA NMR structures.³³ While these calculated pK_a values may identify the correct protonation state to be used in a traditional MD simulation, the latter still lacks the ability to incorporate protonation state information on-the-fly. The ability to perform pH-coupled MD is clearly desirable since it would model realistic pH-dependent responses to structural fluctuations and provide mechanistic insight to RNA-catalyzed reactions.

Received: September 8, 2011

Published: November 22, 2011

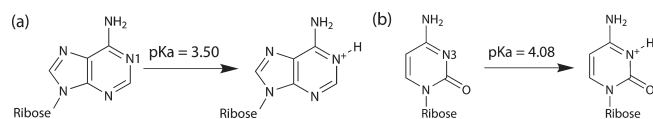


Figure 1. Protonation site of (a) adenosine and (b) cytidine and their respective pK_a values.

In the development of MD simulations, there has been considerable success in calculating pK_a values of protein residues. Warshel and co-workers first demonstrated the feasibility of using microscopic free energy calculations to determine the pK_a values of protein residues.^{34–37} Variations of this approach have been developed that couple the protonation state of a titratable residue with the protein conformation; in these strategies, the atomic coordinates and the protonation state itself evolve according to the dynamics of the system. Two distinct classes of implementation for this methodology exist and differ in the manner in which the titration coordinates are treated—either discretely or continuously. The discrete titration variant is typically implemented by combining MD sampling of the atomic coordinates with Monte Carlo (MC) sampling of protonation states. At regular intervals during a typical MD simulation, a MC step is performed to determine the change of the protonation state. Implementation of discrete constant pH MD (CPHMD) in explicit solvent was first reported by Bürgi et al.³⁸ and Baptista and co-workers,³⁹ and a number of methodological improvements were made by Baptista and co-workers^{40–42} and Stern.⁴³ Discrete CPHMD has also been implemented with implicit solvent by Długosz and Antosiewicz^{44,45} and Mongan et al.,⁴⁶ with improvements to achieve better sampling by Meng et al.⁴⁷ and Williams et al.⁴⁸ More recently, Warshel and co-workers developed a more physically realistic form of pH-dependent MD⁴⁹ based on the time-dependent MC sampling of the proton transfer process,⁵⁰ which uses the empirical valence bond (EVB) framework to simulate a single proton transfer from the protein to a surrounding water molecule.⁵¹ In contrast, in the continuous titration variant of CPHMD, the titration coordinate is propagated continuously between the protonated and unprotonated states. Brooks and co-workers developed CPHMD using implicit solvent^{52–54} which utilized the λ dynamics approach^{55–57} to treat the titration coordinates. Recent work by Shen and co-workers has improved the prediction accuracy of continuous CPHMD,^{58–60} and it has been extended to explicit solvent by Grubmüller and co-workers.⁶¹ Yang and co-workers have also reported using orthogonal space random walk, an enhanced sampling technique compatible with λ -dynamics, that provided accurate pK_a predictions for buried protein residues in explicit solvent simulations.⁶² CPHMD has been used by Brooks and co-workers to investigate numerous pH-dependent conformational changes in proteins,^{63–65} and other investigators in the field have reported similar successes as well.^{66–70} For a more comprehensive overview of CPHMD, we refer our readers to the following review.⁷¹

In this article, we will adopt the newer functional forms of λ developed by Knight and Brooks that have been implemented in multisite λ -dynamics (MS λ D)^{72,73} as the basis of a new MS λ D-based constant pH MD simulation framework (CPHMD^{MS λ D}) and parametrize CPHMD^{MS λ D} to investigate protonation events of nucleic acids in explicit solvent. We will demonstrate the quality of this new CPHMD^{MS λ D} model by its ability to accurately reproduce experimental pK_a values of simple mononucleotide

systems to a mean absolute error of 0.24 pK_a units. To the best of our knowledge, this is the first constant pH MD simulation for nucleic acids to be reported in literature.

2. THEORY

We briefly review the theory behind CPHMD and highlight the relevant modification in our implementation of CPHMD^{MS λ D}. In the original CPHMD model, the protonation/deprotonation process is simulated as a special case of λ -dynamics where the λ variables are used to define titration coordinates.^{55–57} In λ -dynamics, the simulation is under the influence of a hybrid Hamiltonian, and its potential energy is defined by

$$U_{\text{tot}}(X, \{x\}, \{\lambda\}) = U_{\text{env}}(X) + \sum_{\alpha=1}^{N_{\text{sites}}} [\lambda_{\alpha,1}(U(X, x_{\alpha,1})) + \lambda_{\alpha,2}(U(X, x_{\alpha,2}))] \quad (1)$$

where N_{sites} is the total number of titrating residues, X represents the coordinates of the environment atoms, and $x_{\alpha,1}$ and $x_{\alpha,2}$ represent the coordinates of atoms in residue α that are associated with the protonated and unprotonated states, respectively. The titrating proton and the other atoms whose charges vary according to the protonation state of the residue (usually atoms within 2–3 bonds from the titrating proton) are included in both $x_{\alpha,1}$ and $x_{\alpha,2}$ and are defined as a part of the “titrating fragment.” The scaling factor that is associated with the titrating residue α changes dynamically throughout the simulation and is described by a set of continuous coordinates that are governed by the following equations:

$$\lambda_{\alpha,1} = \sin^2 \theta_{\alpha} \text{ and } \lambda_{\alpha,2} = 1 - \sin^2 \theta_{\alpha} \quad (2)$$

The end points define the physically relevant protonated ($\lambda_{\alpha,1} = 1$, $\lambda_{\alpha,2} = 0$) and unprotonated ($\lambda_{\alpha,1} = 0$, $\lambda_{\alpha,2} = 1$) states. In recent work, Knight and Brooks developed the alternative λ^{Nexp} functional form of λ :

$$\lambda_{\alpha,i}^{\text{Nexp}} = \frac{e^{\text{csin } \theta_{\alpha,i}}}{\sum_{j=1}^N e^{\text{csin } \theta_{\alpha,j}}} \quad (3)$$

When applied to the two-state system representing the protonated and unprotonated forms this functional form becomes

$$\lambda_{\alpha,1} = \frac{e^{\text{csin } \theta_{\alpha,1}}}{e^{\text{csin } \theta_{\alpha,1}} + e^{\text{csin } \theta_{\alpha,2}}} \text{ and } \lambda_{\alpha,2} = \frac{e^{\text{csin } \theta_{\alpha,2}}}{e^{\text{csin } \theta_{\alpha,1}} + e^{\text{csin } \theta_{\alpha,2}}} \quad (4)$$

This new form implicitly satisfies the constraints as required by λ -dynamics:

$$0 \leq \lambda_i \leq 1 \text{ and } \sum_{i=1}^N \lambda_i = 1 \quad (5)$$

The use of the λ^{Nexp} functional form also expands the future functionality of our CPHMD^{MS λ D} model to titrate between more than two states, such as the tautomeric forms of nucleic acids.

In CPHMD simulations, the overall free energy of deprotonation of a given residue, $\Delta G_{\text{exp}}(\text{RNA})$, is obtained by calculating the difference between the free energy of deprotonation in the RNA environment, $\Delta G_{\text{sim}}(\text{RNA})$, compared to that of a model compound in solvent, $\Delta G_{\text{sim}}(\text{model})$. By equating this difference

of free energies between the simulated system to that of the experimental system, we obtain

$$\Delta G_{\text{exp}}(\text{RNA}) - \Delta G_{\text{exp}}(\text{model}) = \Delta G_{\text{sim}}(\text{RNA}) - \Delta G_{\text{sim}}(\text{model}) \quad (6)$$

which can be rearranged to estimate the experimental free energy of deprotonation of the RNA:

$$\Delta G_{\text{exp}}(\text{RNA}) = \Delta G_{\text{sim}}(\text{RNA}) - \Delta G_{\text{sim}}(\text{model}) + \Delta G_{\text{exp}}(\text{model}) \quad (7)$$

The free energy of deprotonation of the model compound may also be expressed as

$$\Delta G_{\text{sim}}(\text{model}) = \ln(10)k_{\text{B}}T(\text{p}K_{\text{a}} - \text{pH}) \quad (8)$$

From this perspective, titratable groups in the RNA can be viewed as model compounds that are perturbed by the introduction of the RNA environment via nonbonded interactions, and this is the fundamental expression that needs to be calibrated for each titratable residue in our model, in the present study, adenosine and cytidine. For the initial calibration, the free energy of deprotonation of each isolated model compound calculated using traditional λ -dynamics provided the $\Delta G_{\text{sim}}(\text{model})$ value. When $\Delta G_{\text{sim}}(\text{model})$ is applied to the simulation as a bias, it results in a zero free energy difference between the protonated and unprotonated states, and this condition is equivalent to $\text{pH} = \text{p}K_{\text{a}}$. To simulate the system under different pH environments, eq 8 is used to derive the equivalent $\Delta G_{\text{sim}}(\text{model})$ value that should be applied to the simulation. The reference $\text{p}K_{\text{a}}$ value used in eq 8 was obtained from experimental $\text{p}K_{\text{a}}$ values that were measured at zero ionic strength.¹

In our implementation of CPHMD^{MS λ D}, two biases (F^{fixed} and F^{var}) are incorporated into the potential energy function, and the resulting total potential energy function in our CPHMD simulation may be written as

$$U_{\text{tot}}(X, \{x\}, \{\lambda\}) = U_{\text{env}}(X) + \sum_{\alpha=1}^{N_{\text{sites}}} [\lambda_{\alpha,1}(U(X, x_{\alpha,1}) - F_{\alpha,1}^{\text{fixed}}) + \lambda_{\alpha,2}(U(X, x_{\alpha,2}) - F_{\alpha,2}^{\text{fixed}}) + F_{\alpha,1}^{\text{var}}(\lambda_{\alpha,1}) + F_{\alpha,2}^{\text{var}}(\lambda_{\alpha,2})] \quad (9)$$

In this formalism, the fixed biasing potential that is applied to the unprotonated state ($F_{\alpha,2}^{\text{fixed}}$) represents the calibrated $\Delta G_{\text{sim}}(\text{model})$ value. The other fixed biasing potential applied to the protonated state ($F_{\alpha,1}^{\text{fixed}}$) is kept at zero. Using this setup, when the titration coordinates are allowed to propagate dynamically, the two end points that correspond to physical states may not be well-sampled. Thus, we included the variable biasing potential (F^{var}) which applies an additional bias to encourage sampling of physical states. Identical variable biases are applied to both protonation states.

$$F_{\alpha,i}^{\text{var}} = \begin{cases} k_{\text{bias}}(\lambda_{\alpha,i} - 0.8)^2; & \text{if } \lambda_i < 0.8 \\ 0; & \text{otherwise} \end{cases} \quad (10)$$

The populations of unprotonated (N^{unprot}) and protonated (N^{prot}) states are extracted from the λ trajectory at each pH value, which are used to derive the unprotonated fraction (S^{unprot}):

$$S^{\text{unprot}}(\text{pH}) = \frac{N^{\text{unprot}}(\text{pH})}{N^{\text{unprot}}(\text{pH}) + N^{\text{prot}}(\text{pH})} \quad (11)$$

Overall $\text{p}K_{\text{a}}$ values can be calculated by running simulations at different pH values and fitting the series of S^{unprot} values to a more generalized version of the Henderson–Hasselbach formula:

$$S^{\text{unprot}}(\text{pH}) = \frac{1}{1 + 10^{-n(\text{pH} - \text{p}K_{\text{a}})}} \quad (12)$$

where n is the Hill coefficient. In this formalism, n has a theoretical value of one, and deviations from this value indicate the degree of cooperativity ($n > 1$) or anticooperativity ($n < 1$) between strongly interacting titratable groups.^{74,75}

3. METHODS

3.1. Generating Input Structures. The input structures of the nucleic acids that were used for the simulations were generated from CHARMM topology files using the IC facility in CHARMM, while hydrogen atoms were added using the HBUILD facility.⁷⁶ Model compounds, adenosine and cytidine, were solvated in a cubic box of explicit TIP3P water molecules⁷⁷ of length ~ 20 Å using the convpdb.pl tool from the MMTSB toolset.⁷⁸ The test compounds, adenosine monophosphate (AMP) and cytidine monophosphate (CMP) and dinucleotide sequences of CYT-CYT, ADE-ADE, and CYT-ADE were solvated in a cubic box of explicit water molecules of length ~ 50 Å using the convpdb.pl tool from the MMTSB toolset. The ionic strength was simulated by adding the appropriate number of Na^+ and Cl^- ions to match experimentally reported salt concentrations using convpdb.pl. For the mononucleotides, two isomers in the form of 5'-phosphate and 3'-phosphate were constructed using the patch keywords SPHO and 3PHO, respectively, in CHARMM. All other nucleic acid structures had hydroxyl groups patched to the terminal ends via patch keywords 5TER and 3TER. Additional patches were constructed to represent the protonated forms of adenine and cytosine, and all of the associated bonds, angles and dihedrals were explicitly defined in the patch. Each titratable residue was simulated as a hybrid model that explicitly included atomic components of both the protonated and unprotonated forms. The titratable fragment included the nitrogen atom that is protonated, the protonated hydrogen, and adjacent atoms whose partial charge differed according to the protonation state (see Table 1 and corresponding Tables S1–S3 in the Supporting Information). The environment atoms were defined as all atoms that were not included in the titratable fragments.

3.2. MD Simulations. MD simulations were performed within the CHARMM macromolecular modeling program (version c36a6)⁷⁶ using the CHARMM36 all-atom force field for RNA and TIP3P water.⁷⁹ The simulation set up for λ dynamics is similar to that reported by Knight and Brooks.^{72,73} The SHAKE algorithm⁸⁰ was used to constrain the hydrogen heavy-atom bond lengths. The Leapfrog Verlet integrator was used with an integration time step of 2 fs. A nonbonded cutoff of 15 Å was used with an electrostatic force shifting function, and a vdW switching function between 10 Å and 12 Å. The λ dynamics was performed within the BLOCK facility using the MS λ D framework (MSLD) and selecting the λ^{Nexp} functional form for λ (FNEX). Linear scaling by λ was applied to all energy terms except bond, angle, and dihedral terms, which were treated at full strength regardless of λ value to retain physically reasonable geometries. Each θ_{α} was assigned a fictitious mass of $12 \text{ amu} \cdot \text{Å}^2$, and λ values were saved every 10 steps. The threshold value for assigning $\lambda_{\alpha,i} = 1$ was $\lambda_{\alpha,i} \geq 0.8$. Variable biases (F^{var}) were added

Table 1. Charges and Atom Types Assigned to the Protonated and Unprotonated States of Titratable Nucleic Acids

name	atom	unprotonated		protonated	
		atom type	charge	atom type	charge
ADE	H1	–	–	HN2	0.527
	N1	NN3A	–0.74	NN2G	–0.489
	C2	CN4	0.50	CN4	0.611
	C6	CN2	0.46	CN2	0.571
CYT	H3	–	–	HN2	0.52
	C5	CN3	–0.13	CN3	–0.174
	C2	CN1	0.52	CN1	0.75
	N3	NN3	–0.66	NN2C	–0.874
	C4	CN2	0.65	CN2	0.962
	N4	NN1	–0.75	NN1	–0.654
	H41	HN1	0.37	HN1	0.42
	H42	HN1	0.33	HN1	0.38

to the hybrid potential energy function, and the associated force constant (k_{bias}) was optimized to enhance transition rates between the two protonation states. Since identical k_{bias} values were applied to both protonated and unprotonated states, the PMF at the end-points were not altered, and no reweighting scheme was required. The temperature was maintained at 298 K by coupling to a Langevin heatbath using a frictional coefficient of 10 ps^{-1} . Prior to the simulation, each system was minimized using 300 steps of steepest descents (SD), followed by 200 steps of adopted basis Newton–Raphson. After an initial heating of 4 ps and equilibration of 4 ps, a production run of 1 ns was performed, unless otherwise stated.

3.3. Calculation of pK_a Value. In our protocol, a single S^{unprot} value was estimated by combining the populations of N^{prot} and N^{unprot} from three independent simulations that used different initial seed values. These combined S^{unprot} ratios that were computed at different pH values were then fitted to eq 12 to obtain a single pK_a value. Unless otherwise specified, the reported pK_a value and its error correspond to the mean and standard deviation calculated from three sets of pK_a calculations.

4. RESULTS AND DISCUSSION

4.1. New CHARMM Parameters for Protonated Adenine and Cytosine. We calculated the partial charges for the adenine and cytosine nucleobases in their neutral (unprotonated) and charged (protonated) states using the MMFF94 force field.⁸¹ The change in the partial charge was added to the existing partial charge parameters for neutral adenine and cytosine in CHARMM to assign the charge distribution for the protonated residues. A summary of the differences for charges and atom types between the protonated and unprotonated nucleic acids is reported in Table 1. Parameters for the bond, angle, and dihedral energy terms for the protonated nucleic acid were adapted from existing nucleic acid structures in CHARMM (see Supporting Information). For the protonated adenine, the respective bonded parameters were obtained from guanine, specifically from the six-membered ring component that has atoms analogous to that of adenine (N1, H1, C2, and C6). For the protonated cytosine, the respective bonded parameters were obtained from a tautomeric form of neutral cytosine (obtained from patch CYT1).

4.2. Optimization of Model Potential Parameters. Our CPHMD^{MS λ D} model was implemented using the recently developed $\lambda^{N_{\text{exp}}}$ functional form for λ in MS λ D:⁷³

$$\lambda_{\alpha,i}^{N_{\text{exp}}} = \frac{e^{\text{csin } \theta_{\alpha,i}}}{\sum_{j=1}^N e^{\text{csin } \theta_{\alpha,j}}}$$

Knight and Brooks reported setting the coefficient to 5.5 for the optimal balance between enhancing transition rate and maintaining numerical stability of the integrator in different environments.⁷³ An identical setup was used successfully in our CPHMD^{MS λ D} model. As with the previous implementation of CPHMD for protein residues,^{52–54} we have used the calibrated free energy of deprotonation (G_{bias}) as the fixed biasing potential value in our simulation. The free energy of deprotonation was calibrated for each isolated model compound, i.e., adenosine and cytosine, embedded in explicit solvent using traditional λ -dynamics. In order to facilitate transitions between the two protonation states, we optimized the force constant (k_{bias}) on the variable biasing potential that was applied for each model compound.

It is interesting to note that without the application of the variable bias, no transitions between the protonated and unprotonated states were observed at conditions $\text{pH} = \text{pK}_a$, where one should expect equal population of both states and the maximum transition rate between the two states (see Figure 2). At values of $k_{\text{bias}} < 20 \text{ kcal/mol}$, there were very few transitions in λ phase space between the two states for the entire duration of a 1 ns trajectory. At values $k_{\text{bias}} > 40 \text{ kcal/mol}$, transitions were rapid, but the end states were not adequately sampled. The optimal value of k_{bias} for each nucleoside was selected by considering the competing needs for a high number of transitions and an adequate sampling of the end points (i.e., maintaining a high fraction of physical ligands (FPL) that were sampled). As illustrated in Figure 3, these two properties were observed to be anticorrelated to each other, and there is a distinct range of k_{bias} values (between 25 and 35 kcal/mol) that yielded good transition rates, and where more than 80% of the simulation is spent at the physically relevant end points. The optimized parameters for the two model potentials are reported in Table 2.

The variable bias with a relatively large force constant of 28–30 kcal/mol that is required to achieve a reasonable number of transitions in our simulation may be rationalized by noting that the appearance of a full charge unit when titrating between the two states is likely to significantly perturb the solvent environment around the nitrogen atom. We suggest that time is required for the solvent to reorganize and fully accommodate the new charge distribution as the system titrates from the unprotonated to the protonated state. Figure 4 provides a comparison of the radial distribution function (RDF) of water molecules surrounding the N1 atom of adenosine in its protonated and unprotonated state and indicates that considerable rearrangement of the first solvent shell upon ionization of the residue does occur. For the RDF that describes the distances between N1 and the TIP3P oxygen atoms, we observed that the charged protonated state had a first solvation shell (2.7 Å) that is slightly closer than the uncharged unprotonated state (2.9 Å). A more significant change, however, was observed for the RDF that describes the distances between N1 and the TIP3P hydrogen atoms in which the protonated state first solvation shell was pushed back (3.4 Å) compared to that of the unprotonated state (2.0 Å). These observations are consistent with the expectation that water

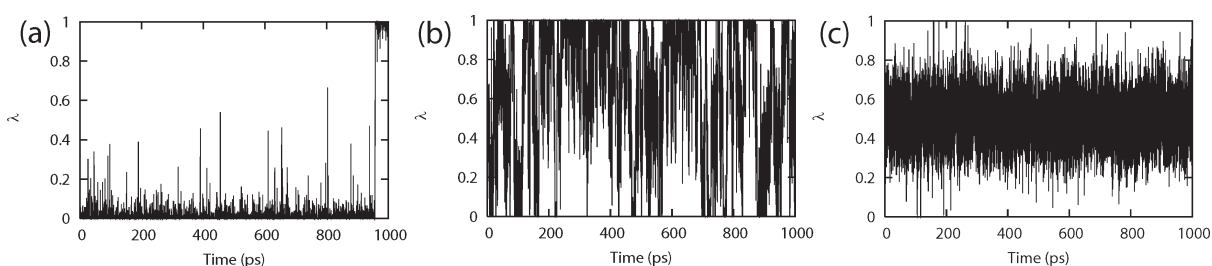


Figure 2. Transitions between the protonated and unprotonated state of adenosine in λ phase space at $\text{pH} = \text{p}K_a$ for a 1 ns trajectory with varying k_{bias} values of (a) 20, (b) 30 and (c) 40.

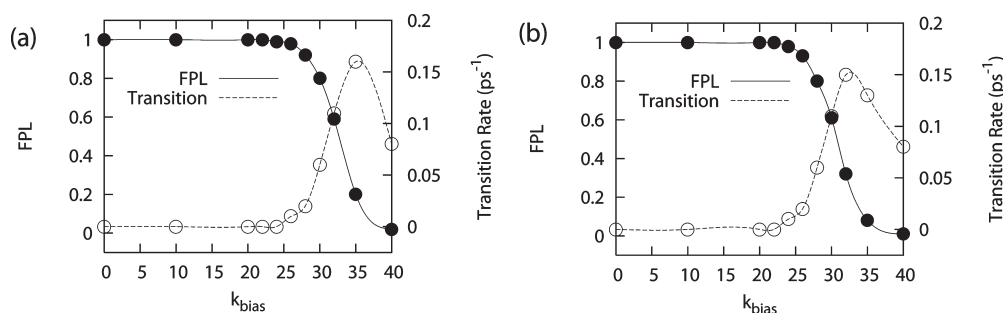


Figure 3. Effect of increasing k_{bias} on the transition rate and FPL for (a) adenosine and (b) cytosine. Sampling characteristics were obtained from 5 independent MD runs of 1 ns each.

Table 2. Parameters for the Model Potential^a

nucleotide	G_{bias} (kcal/mol)	k_{bias} (kcal/mol)	reference $\text{p}K_a$ ^b
adenine	19.39	29.75	3.50
cytosine	75.24	27.75	4.08

^a G_{bias} was assigned to be the free energy of deprotonation of adenosine or cytosine at zero ionic strength, and k_{bias} was optimized to achieve a maximum transition rate while maintaining physical states for more than 80% of the entire trajectory. ^b Reference $\text{p}K_a$ is the experimental $\text{p}K_a$ values for the model compounds that were measured at zero ionic strength.¹

molecules would orient their hydrogen atoms toward the partial negative charge of the nitrogen atom in the unprotonated state and subsequently would flip their hydrogen atoms “outwards” and orient their oxygen atoms closer toward the partial positive charge of the protonated hydrogen that is present in the protonated state. Similar trends were observed for the RDF of water molecules that surround the N3 atom of cytidine (data not shown). An analogous change in RDF of water molecules around the protonated N5 atom of the substrate of dihydrofolate reductase was also observed with MD simulations that sampled different protein conformations that altered the water accessibility of the ligand pocket.⁸²

4.3. Sampling Efficiency of Explicit Solvent CPHMD Simulations. The sampling efficiency as measured by the transition rates between the two protonation states in our CPHMD^{MS λ D} model is quite good with ~ 50 transitions per ns for our model compounds at $\text{pH} = \text{p}K_a$. Given that the solvent reorganization upon the perturbation of a full charge unit was reported to be on a time scale of up to 3 ps in previous MD simulations³⁹ and that the mean duration of the physically relevant protonation states in our simulations is 20 ps, the sampling characteristics of our system

are sufficient to allow solvent reorganization to occur. However, the transition rate is markedly lower than what has been observed in CPHMD simulations that are performed using implicit solvent models.^{52,53} It should be noted that our model potential parameters, specifically the k_{bias} values as implemented in CPHMD^{MS λ D} have been selected conservatively. For example, the transition rate can be doubled at the expense of reducing the FPL to 0.6 (Figure 3) which, provided that simulations are long enough to sufficiently enumerate the relative end-state populations, may be a better option for simulating full RNA systems where observing transitions between protonation states may be more challenging.

The more limited sampling efficiency of explicit solvent CPHMD simulations was also recently reported by Grubmüller and co-workers where the titration of an imidazole model compound achieved ~ 100 transitions in a 20 ns trajectory,⁶¹ which is a rate of ~ 5 transitions per ns. Considering the computational expense of performing explicit solvent simulations, our rate of ~ 50 transitions per ns that is achieved with the optimization of our implementation of explicit solvent CPHMD is clearly advantageous. Finally, in Table 3, we present a comparison between the sampling characteristics of our simulation to that of previous work performed in the MS λ D framework by Knight and Brooks for modeling series of inhibitors of HIV-1 reverse transcriptase.⁷² Using the same force constant for the variable bias (i.e., $k_{\text{bias}} = 7$) as what was previously reported, we observed a significant drop in sampling performance with virtually no transitions observed between the two protonation states at $\text{pH} = \text{p}K_a$. Our optimization of k_{bias} assisted in improving the sampling characteristics, but the transition rate still remains about 4-fold less efficient than previous work. We note that earlier work performed by Knight and Brooks modeled hybrid ligands in which the substituents did not differ significantly in

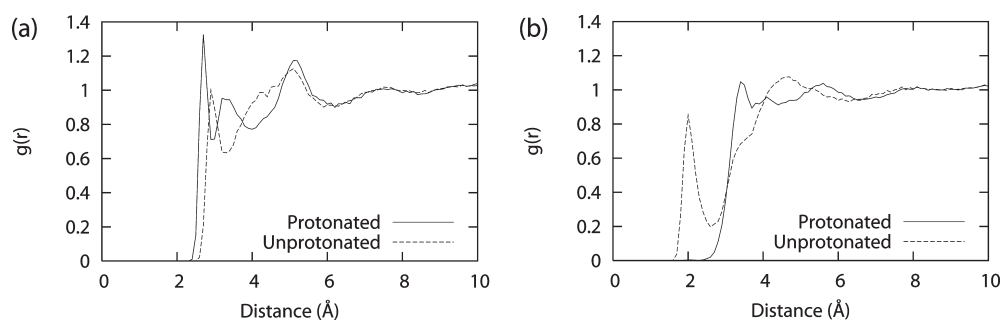


Figure 4. RDF of water molecules for (a) N1(ADE)-O(TIP3P) distances and (b) N1(ADE)-H(TIP3P) distances within a sphere of 10 Å from the N1 atom of adenosine in both protonated and unprotonated states.

Table 3. Sampling Characteristics of Simulations Performed at pH = pK_a

	previous work ^a	adenosine (default)	cytidine (default)	adenosine (optimized)	cytidine (optimized)
k_{bias}	7.00	7.00	7.00	29.75	27.75
FPL	0.780	1.000	1.000	0.828	0.832
transitions (ps ⁻¹)	0.190	0.001	0.001	0.050	0.051

^a Sampling characteristics of a two-state hybrid ligand in explicit water investigated in previous work (obtained from Table 3, hybrid ligand F).⁷²

terms of their partial charge distributions. Thus, the introduction of a full charge unit when titrating between the two states in CPHMD^{MSλD} is likely to be the primary cause for the reduction sampling efficiency that we observe in the present simulations.

4.4. Convergence and Precision of Calculations. The challenges associated with sampling and convergence for CPHMD simulations have been reported on several occasions,^{48,54} and these are expected to be an even greater concern in explicit solvent CPHMD where sampling efficiency is reduced. To validate the robustness of our CPHMD^{MSλD} model in its ability to achieve adequate convergence, we performed a series of simulations at pH = pK_a for our model compounds. The degree of convergence in our simulations was determined by calculating the unsigned deviation between the free energy of protonation, estimated from subsets of shortened trajectories, and the free energy of protonation that was estimated from 10 1 ns trajectories. Different combinations of trajectory length and number of independent runs were systematically examined to determine the most cost-effective trade-off between computational expense and precision of the calculations. The results are summarized in Figure 5. It was observed that individual trajectories required at least 100 ps to reliably observe any transitions between protonation states. In fact, we observed that a minimum simulation time of ~500 ps per trajectory was required to obtain a precision of ~0.20 kcal/mol in our calculations (Figure 5a), and running multiple shorter independent runs would not produce converged results unless the 500 ps threshold was crossed. Our results indicate that good precision can be achieved by using a total simulation time of 3 ns in the form of 3 independent runs of 1 ns each, where the unsigned deviations for the free energy of deprotonation was 0.05 kcal/mol for adenosine (Figure 5b). It should be noted, however, that this level of precision was achieved in previous work three times more quickly for hybrid ligands whose charge distributions were similar. All subsequent calculations of pK_a values in this paper were estimated using three independent runs of 1 ns each.

The performance of the multisite λ-dynamics (MSλD) approach, on which CPHMD^{MSλD} is based, has been evaluated in comparison to traditional FEP calculations by Knight and Brooks.⁷² For substituents with similar charge distributions, it has been established that relative hydration and relative binding free energies calculated from both MSλD and FEP are in good agreement with each other and that MSλD was three times more efficient than regular FEP. Our current work involves substituents that have significantly different charge distributions from one another, i.e., the charges associated with the protonated and deprotonated states, respectively, and consequently CPHMD^{MSλD} takes longer to converge. Analogously, we expect that FEP calculations will also take longer than what was reported in Knight and Brooks but would still be less efficient in their convergence than MSλD calculations. In addition, traditional FEP calculations are not well-suited for simultaneously exploring multiple titrating sites or multiple tautomers, and so the MSλD-based approach of CPHMD^{MSλD} is more generalizable than FEP methods to model these more complex situations.

4.5. Calibration Curve of Model Systems: Adenosine and Cytidine. We calibrated our CPHMD^{MSλD} model at 298 K using zero salt concentration. The reference pK_a that was used in the calibration was the experimental pK_a that was measured under similar conditions (25 °C at zero ionic strength).¹ The titration curve of the model nucleoside compounds, adenosine and cytidine, is shown in Figure 6. The best-fit Henderson–Hasselbalch curve has a near ideal Hill coefficient for adenosine ($n = 0.94$) and cytidine ($n = 0.93$). The calculated pK_a value of 3.50 for adenosine was in excellent agreement with experimental values, and the pK_a of 4.22 for cytidine is only slightly higher than the reference value by 0.14 pK_a units. The accuracy of the calculated pK_a values is determined primarily by the sampling efficiency at pH = pK_a and the quality of the calibration of the G_{bias} values that are used to simulate distinct pH conditions. Our results demonstrate that a series of 3 × 1 ns simulations is sufficient to provide reasonably accurate results, which is significantly less than the 20 ns trajectory employed by Grubmüller and co-workers in their explicit solvent CPHMD model.⁶¹

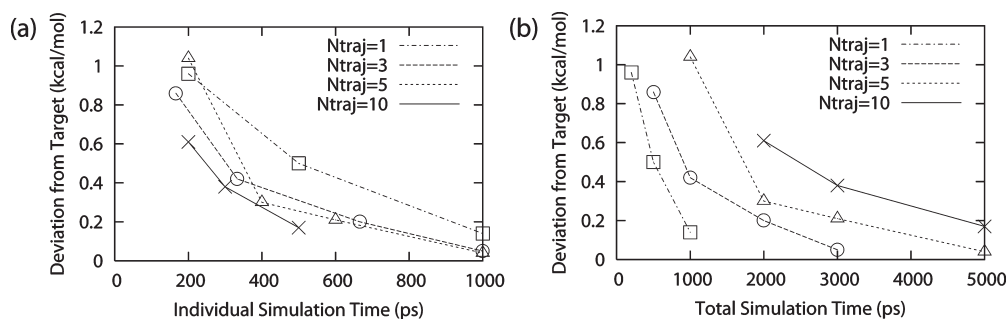


Figure 5. Unsigned deviation for the free energy of deprotonation of adenosine as a function of (a) total simulation time from all N trajectories and (b) individual simulation time of each of the N trajectories.

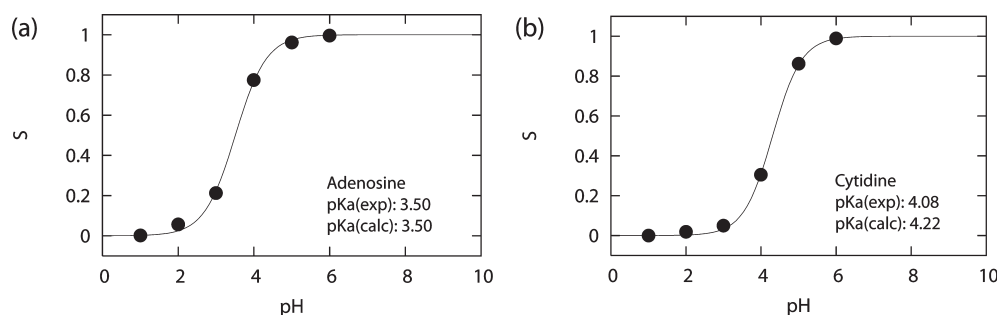


Figure 6. Sample titration curves for model nucleoside compounds (a) adenosine and (b) cytidine.

Table 4. Calculated and Experimental pK_a Values of Test Compounds

compound	[NaCl] (M)	calculated	experimental	abs. error
β -AMP-3	no salt	4.20 ± 0.06	—	—
β -AMP-3	0.15	3.79 ± 0.11	3.65	0.14
AMP-5	no salt	4.08 ± 0.03	—	—
AMP-5	0.15	3.89 ± 0.16	3.74	0.15
β -CMP-3	no salt	4.77 ± 0.05	—	—
β -CMP-3	0.15	4.56 ± 0.10	4.31	0.25
CMP-5	no salt	4.90 ± 0.07	—	—
CMP-5	0.10	4.67 ± 0.08	4.24	0.43

4.6. Quantitative pK_a Value Calculations for Simple Nucleotides. First, we tested our CPHMD^{MS λ D} model on single nucleotide test compounds, adenine monophosphate (AMP) and cytosine monophosphate (CMP), at zero ionic strength, and the results are summarized in Table 4. The calculated pK_a values for AMP-5 and β -AMP-3 were 4.08 and 4.20, respectively. Compared to adenosine, the pK_a values of these nucleotide counterparts were slightly elevated by ~ 0.5 pK_a units. Similarly, the nucleotide counterparts of cytidine with pK_a values for CMP-5 and β -CMP-3 of 4.90 and 4.77, respectively, had slightly elevated pK_a values by ~ 0.5 pK_a units compared to cytidine. The calculated pK_a values for both 5'- and 3'-phosphate isomers of adenosine and cytosine are not statistically different at the 95% confidence interval. The increase in the calculated pK_a values from their nucleoside counterparts is expected, since the presence of the negative charge from the phosphate group may interact with the positively charged protonated base and weakly stabilize it, thus increasing the population of the protonated state and causing a corresponding increase in the calculated pK_a value.

In order to compare our calculated pK_a values with experimental results, we performed simulations that mimicked the ionic strength of the environment (i.e., 100–150 mM NaCl) in which the experiments were performed.^{83,84} By explicitly incorporating the salt environment, the calculated pK_a values are systematically lowered relative to those obtained from the zero ionic strength simulations. This shift in pK_a values is to be expected since the presence of Na^+ ions screens the electrostatic effects of the phosphate group. The results in Table 4 indicate that our pK_a predictions had an average absolute error of 0.24 pK_a units compared to experiment, and we conclude that our CPHMD^{MS λ D} model is capable of making accurate quantitative predictions of pK_a values for simple nucleotides. These results also indicate that our model is capable of accounting for the differences between zero and nonzero ionic strength environments and highlights the importance of simulating the system at the appropriate ionic strength to mimic experimental conditions.

4.7. Modeling Interactions between Adjacent Titrating Residues. Finally, we tested our CPHMD^{MS λ D} model on dinucleotide sequences ADE-ADE, CYT-CYT, and CYT-ADE at zero ionic strength, where both nucleotides were titrated simultaneously in the same simulation. The pK_a values were shifted upward compared to the nucleoside model compounds for all sequences, ADE-ADE (4.08 ± 0.20 and 4.06 ± 0.16), CYT-CYT (4.93 ± 0.05 and 4.76 ± 0.09), and CYT-ADE (5.06 ± 0.07 and 3.85 ± 0.26) and were similar to the corresponding mononucleotide pK_a values. For some of the sets of pK_a calculations for the dinucleotide sequences, the Hill coefficient had more significant deviations from one compared to the monomeric compounds. Specifically, the value was lowered ($n < 0.8$) for 5 of the 9 sets of pK_a calculations. When the Hill coefficient deviates from one, it suggests that adjacent residues

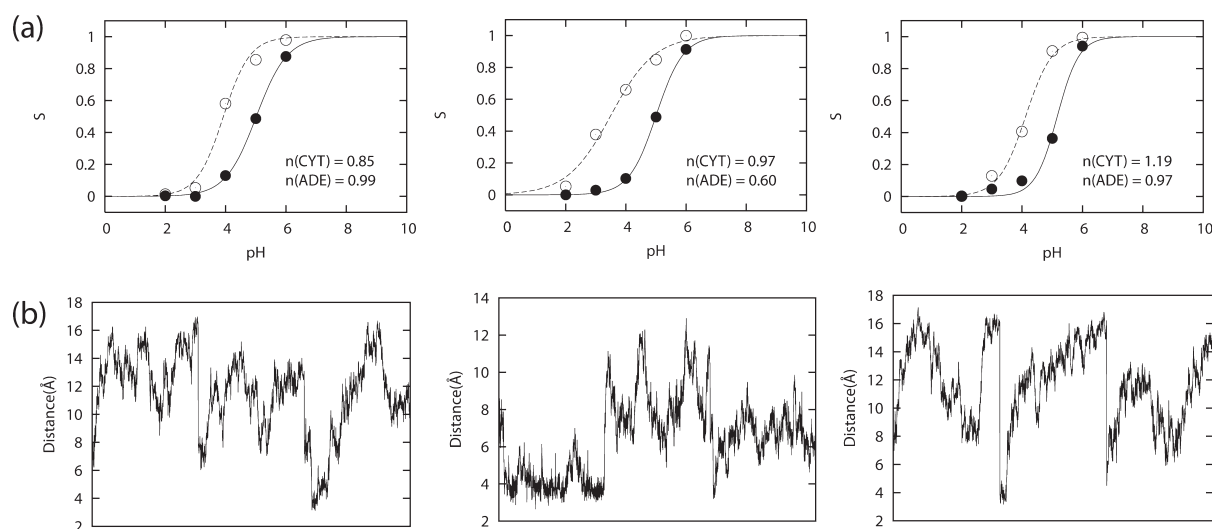


Figure 7. (a) Titration curves for CYT-ADE and (b) time series of distance between N3 CYT and N1 ADE atoms at pH 3 for all 3 sets of pK_a calculations.

are interacting with each other in either a cooperative ($n > 1$) or an anticooperative ($n < 1$) fashion. Cross-correlation analysis of the protonation states (data not shown) however indicates only weakly correlated behavior, which suggests that the interaction between adjacent residues is not strong. The second set of pK_a calculations on CYT-ADE exhibited the lowest Hill coefficient ($n = 0.60$), indicating the strongest anticooperative behavior. Analysis of the individual titration curves as shown in Figure 7 indicates that the S^{unprot} ratio shows the greatest deviation between the second set and the other two sets at pH 3. We analyzed the mean distance between the nitrogen atom that is protonated in CPHMD (i.e., N3 CYT and N1 ADE) of adjacent residues at pH 3, and the results are shown in Figure 7. In one simulation of the second set, the mean distance sampled was about 4–6 Å, in comparison to the typical values of 8–16 Å for all other simulations. We suggest that this simulation contributed significantly to the higher S^{unprot} ratio for the second set that in turn gave rise to the lower Hill coefficient. The lack of strong interactions between adjacent titrating residues in the other two sets of pK_a calculations of CYT-ADE is apparently due to the lack of sampling of configuration space in which these two residues are close enough to influence each other's protonation state. We suggest that stronger cooperative or anticooperative effects are likely to be observed when modeling RNA structures with stable conformations in which the nucleobases are held in close proximity to one another.

4.8. Moving toward CPHMD of Full RNA Systems. The remarkable agreement of our calculated pK_a values to experiment is encouraging; however, several challenges may be anticipated when applying our CPHMD^{MSAD} model to full RNA structures.

First, instead of isolated monomeric compounds, the titratable residues of interest in RNA macromolecules are nucleotides that are buried in the interior of the RNA which interact with multiple residues (e.g., via base-pairing interactions). Therefore, the increased perturbation of the titrating residue's local environment and the increased complexity arising from nonbonded interactions with adjacent residues are likely to reduce sampling efficiency in full RNA structures. We illustrate this claim with a hypothetical example of a residue whose pK_a value varies with RNA conformation. The formation of a Watson–Crick A–T base

pair involves the N1 atom of adenine, and this results in a depression of its pK_a value relative to the isolated base since the nitrogen atom (that would be protonated in CPHMD) is now serving as a hydrogen-bond acceptor. Conformational fluctuations, such as base flipping motions, may expose these buried nucleotides to the solvent and cause a corresponding increase in their respective pK_a values. In order to reproduce experimentally measured pK_a values from NMR studies in the above example, it may be necessary to sample these two conformational states that are observed at the time scale of which the measurements were taken. The use of advanced sampling methods, such as replica exchange, has been previously implemented in protein CPHMD by Brooks and co-workers to improve sampling performance⁵⁴ and when used with CPHMD^{MSAD} may be expected to yield similar improvements in model performance. Similarly, accelerated MD⁸⁵ has been implemented with CPHMD, and it has yielded improvements in conformation sampling.⁴⁸ Other sampling methods developed to sample long time scale conformational changes, such as self-guided langevin dynamics,^{86,87} may also achieve a similar effect.

Another challenge may arise from the sampling in λ phase space in the presence of many interacting titratable groups. Under the physiologically relevant pH range, protein residues typically have 4 titratable residues out of 20 amino acids. In contrast, half of the nucleic acid building blocks are titratable in our current CPHMD^{MSAD} model. While it may be common that titratable residues on a protein are separated from one another in terms of spatial proximity, and thus the state of one titratable residue is unlikely to influence the others, this is not the case for RNA. In the absence of prior information about which residues will likely be in which protonation states, all adenosine and cytosine residues in the macromolecule could be modeled as titratable. However, in this case, the cooperative or anticooperative effect that these simultaneously titrating residues have on one another could be significant. Such a situation may lead to hidden barriers in adjacent λ phase spaces, in which the λ values of residue i are restricting the propagation of the λ values of residue j . Thus, the efficiency of sampling in λ phase space would need to be improved, and a lower FPL without compromising the physical accuracy of the model may be necessary in order to

maintain a reasonable minimum transition rate between the two protonation states. A recently developed enhanced sampling technique, orthogonal space random walk (OSRW) has been developed to address such sampling challenges associated with strongly coupled hidden free energy barriers, and it has been successfully applied to predict the pK_a values of buried protein residues that are typically not accurately reproduced in conventional CPHMD approaches. The implementation of OSRW with CPHMD^{MS λ D} could potentially model strongly coupled titrating residues with better accuracy.^{62,88}

For CPHMD^{MS λ D} to successfully investigate pH-dependent properties of RNA structures over a longer time scale (μ s and beyond) or pH-dependent properties of large RNA structures, such as a ribosomal subunit, a reduction of the computational expense that is associated with explicit solvent CPHMD is desirable. Greater computational efficiency may be achieved by using hybrid explicit/implicit solvation models, in which a few layers of explicit solvent water molecules are placed near the RNA surface with the rest of the environment described by an implicit solvation model.^{89–91} Other models reduce the number of explicit waters in the simulation by using a thin shell of explicit water and hold these waters near the RNA surface with a restraining force.^{92,93} The use of the surface constrained all atom solvent model that requires fewer explicit water molecules than the periodic boundary conditions implementation in MD simulations^{94,95} has also been validated on a number of pK_a calculations performed on protein residues.^{37,96} Multiscale modeling approaches that use a coarse-grained model to provide the reference potential for the thermodynamic cycle that can be used to speed up the free energy calculations of protonation events to simulate pH-dependent dynamics have also been reported recently.^{49,97} Alternatively, we have seen considerable advances in implicit solvent models in recent years,⁹⁸ and they have been successfully implemented with protein CPHMD.^{52–54} Established work in parametrizing implicit solvation models for RNA is encouraging,⁹⁹ but ongoing work in our lab (unpublished results) indicates that implicit solvent models do not accurately reproduce explicit solvent simulation results when simulating RNA. Therefore, the successful implementation of an implicit solvation model to CPHMD^{MS λ D} would be an avenue for future development.

Finally, while the use of NaCl may serve to reproduce the ionic strength environment at which experimental studies are conducted, divalent ions, such as Mg^{2+} , play functional roles in many RNA structures, and current parameters would need to be examined to ensure their ability to reproduce experimental observables in RNA macromolecules. Our CPHMD^{MS λ D} model may also be expanded to include titratable residues of guanine, uracil, and thymine. Although the bulk of experimental studies have implicated adenine and cytosine as key protonated residues in RNA, there is some evidence that suggests that the presence of protonated guanine, such as the G8 residue in the active site of the hairpin ribozyme, may also play a mechanistic role.¹⁰⁰ The adoption of the $\lambda^{N_{exp}}$ functional form for λ in our CPHMD^{MS λ D} model also allows us to expand the representation of the titratable fragments to include tautomeric forms of nucleotides in both the unprotonated and protonated states. Recent theoretical studies have also suggested that stable tautomers may exist under specific conditions.^{101–103} Thus, the ability to titrate among four states (unprotonated, unprotonated tautomer, protonated and protonated tautomer) may assist in clarifying the structural or mechanistic roles that involve tautomeric forms in RNA.

5. CONCLUSION

In conclusion, we have parametrized a protonated adenine and protonated cytosine for use in the first reported constant pH molecular dynamics (CPHMD) for nucleic acids. We have adopted the new functional form $\lambda^{N_{exp}}$ for λ that was recently developed for multisite λ -dynamics (MS λ D) and demonstrate good sampling characteristics in which rapid and frequent transitions between the protonated and unprotonated states at $pH = pK_a$ are achieved, while sampling the physically relevant protonation states for more than 80% of the trajectory. Compared to existing implementations of explicit solvent CPHMD, the sampling in our method sees a 10-fold improvement, while maintaining sufficient residency time of the physical protonation states to ensure proper solvent reorganization. A series of 3 independent runs of 1 ns each was determined to be sufficiently precise for calculating the pK_a values for simple nucleotide systems. pK_a values calculated for simple nucleotides are in a good agreement with experimentally measured values with a mean absolute error of 0.24 pK_a units, affirming that our CPHMD^{MS λ D} model has the ability to make accurate quantitative predictions for simple nucleotide systems. Our work paves the way for the deployment of CPHMD as a powerful tool to investigate pH-dependent biological properties of RNA macromolecules.

■ ASSOCIATED CONTENT

S Supporting Information. Parameters for the bond, angle, and dihedral energy terms for protonated adenine and cytosine. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: brookscl@umich.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENT

This work was supported by grants from the National Institutes of Health (GM037554 and GM057053).

■ REFERENCES

- (1) Izatt, R. M.; Christensen, J. J.; Rytting, J. H. *Chem. Rev.* **1971**, *71*, 439.
- (2) Gao, X. L.; Patel, D. J. *J. Biol. Chem.* **1987**, *262*, 16973.
- (3) Asensio, J. L.; Lane, A. N.; Dhesi, J.; Bergqvist, S.; Brown, T. *J. Mol. Biol.* **1998**, *275*, 811.
- (4) Jang, S. B.; Hung, L. W.; Chi, Y. I.; Holbrook, E. L.; Carter, R. J.; Holbrook, S. R. *Biochemistry* **1998**, *37*, 11726.
- (5) Bink, H. H.; Hellendoorn, K.; van der Meulen, J.; Pleij, C. W. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 13465.
- (6) Morse, S. E.; Draper, D. E. *Nucleic Acids Res.* **1995**, *23*, 302.
- (7) Leontis, N. B.; Stombaugh, J.; Westhof, E. *Nucleic Acids Res.* **2002**, *30*, 3497.
- (8) Lee, J. C.; Gutell, R. R. *J. Mol. Biol.* **2004**, *344*, 1225.
- (9) Durant, P. C.; Davis, D. R. *J. Mol. Biol.* **1999**, *285*, 115.
- (10) Chen, G.; Kennedy, S. D.; Turner, D. H. *Biochemistry* **2009**, *48*, 5738.
- (11) Muth, G. W.; Chen, L.; Kosek, A. B.; Strobel, S. A. *RNA* **2001**, *7*, 1403.

- (12) Siegfried, N. A.; O'Hare, B.; Bevilacqua, P. C. *Biochemistry* **2010**, *49*, 3225.
- (13) Kim, M.; Huang, T.; Miller, J. H. *J. Bacteriol.* **2003**, *185*, 4626.
- (14) Giri, I.; Stone, M. P. *Biochemistry* **2003**, *42*, 7023.
- (15) Nakano, S.; Chadalavada, D. M.; Bevilacqua, P. C. *Science* **2000**, *287*, 1493.
- (16) Bevilacqua, P. C.; Brown, T. S.; Nakano, S.; Yajima, R. *Biopolymers* **2004**, *73*, 90.
- (17) Das, S. R.; Piccirilli, J. A. *Nat. Chem. Biol.* **2005**, *1*, 45.
- (18) Wilson, T. J.; Ouellet, J.; Zhao, Z. Y.; Harusawa, S.; Araki, L.; Kurihara, T.; Lilley, D. M. *RNA* **2006**, *12*, 980.
- (19) Chen, J.-H.; Yajima, R.; Chadalavada, D. M.; Chase, E.; Bevilacqua, P. C.; Golden, B. L. *Biochemistry* **2010**, *49*, 6508.
- (20) Shih, I. H.; Been, M. D. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 1489.
- (21) Wadkins, T. S.; Shih, I.; Perrotta, A. T.; Been, M. D. *J. Mol. Biol.* **2001**, *305*, 1045.
- (22) Ke, A.; Zhou, K.; Ding, F.; Cate, J. H.; Doudna, J. A. *Nature* **2004**, *429*, 201.
- (23) Gong, B.; Chen, J. H.; Chase, E.; Chadalavada, D. M.; Yajima, R.; Golden, B. L.; Bevilacqua, P. C.; Carey, P. R. *J. Am. Chem. Soc.* **2007**, *129*, 13335.
- (24) Cerrone-Szakal, A. L.; Siegfried, N. A.; Bevilacqua, P. C. *J. Am. Chem. Soc.* **2008**, *130*, 14504.
- (25) Ravindranathan, S.; Butcher, S. E.; Feigon, J. *Biochemistry* **2000**, *39*, 16026.
- (26) Ryder, S. P.; Oyelere, A. K.; Padilla, J. L.; Klostermeier, D.; Millar, D. P.; Strobel, S. A. *RNA* **2001**, *7*, 1454.
- (27) Kuzmin, Y. I.; Da Costa, C. P.; Cottrell, J. W.; Fedor, M. J. *J. Mol. Biol.* **2005**, *349*, 989.
- (28) Nam, K.; Gao, J.; York, D. M. *J. Am. Chem. Soc.* **2008**, *130*, 4680.
- (29) Guo, M.; Spitalo, R. C.; Volpini, R.; Krucinska, J.; Cristalli, G.; Carey, P. R.; Wedekind, J. E. *J. Am. Chem. Soc.* **2009**, *131*, 12908.
- (30) Cottrell, J. W.; Scott, L. G.; Fedor, M. J. *J. Biol. Chem.* **2011**, *286*, 17658.
- (31) Ditzler, M. A.; Sponer, J.; Walter, N. G. *RNA* **2009**, *15*, 560.
- (32) Mlýnský, V.; Banás, P.; Hollas, D.; Réblová, K.; Walter, N. G.; Sponer, J.; Otyepka, M. *J. Phys. Chem. B* **2010**, *114*, 6642.
- (33) Tang, C. L.; Alexov, E.; Pyle, A. M.; Honig, B. *J. Mol. Biol.* **2007**, *366*, 1475.
- (34) Russell, S. T.; Warshel, A. *J. Mol. Biol.* **1985**, *185*, 389.
- (35) Lee, F. S.; Chu, Z. T.; Warshel, A. *J. Comput. Chem.* **1993**, *14*, 161.
- (36) Warshel, A.; Sussman, F.; King, G. *Biochemistry* **1986**, *25*, 8368.
- (37) Sham, Y. Y.; Chu, Z. T.; Warshel, A. *J. Phys. Chem. B* **1997**, *101*, 4458.
- (38) Burgi, R.; Kollman, P. A.; van Gunsteren, W. F. *Proteins* **2002**, *47*, 469.
- (39) Baptista, A. M.; Teixeira, V. H.; Soares, C. M. *J. Chem. Phys.* **2002**, *117*, 4184.
- (40) Machuqueiro, M.; Baptista, A. M. *J. Phys. Chem. B* **2006**, *110*, 2927.
- (41) Baptista, A. M.; Machuqueiro, M. *Proteins* **2008**, *72*, 289.
- (42) Baptista, A. M.; Machuqueiro, M. *J. Am. Chem. Soc.* **2009**, *131*, 12586.
- (43) Stern, H. A. *J. Chem. Phys.* **2007**, *126*, 164112.
- (44) Dlugosz, M.; Antosiewicz, J. M. *Chem. Phys.* **2004**, *302*, 161.
- (45) Dlugosz, M.; Antosiewicz, J. M.; Robertson, A. D. *Phys. Rev. E* **2004**, *69*, 021915.
- (46) Mongan, J.; Case, D. A.; McCammon, J. A. *J. Comput. Chem.* **2004**, *25*, 2038.
- (47) Meng, Y.; Roitberg, A. E. *J. Chem. Theory Comput.* **2010**, *6*, 1401.
- (48) Williams, S. L.; de Oliveira, C. A.; McCammon, J. A. *J. Chem. Theory Comput.* **2010**, *6*, 560.
- (49) Messer, B. M.; Roca, M.; Chu, Z. T.; Vicatos, S.; Kilshtain, A. V.; Warshel, A. *Proteins* **2010**, *78*, 1212.
- (50) Olsson, M. H. M.; Warshel, A. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 6500.
- (51) Aqvist, J.; Warshel, A. *Chem. Rev.* **1993**, *93*, 2523.
- (52) Lee, M. S.; Salsbury, F. R.; Brooks, C. L., III *Proteins* **2004**, *56*, 738.
- (53) Khandogin, J.; Brooks, C. L., III *Biophys. J.* **2005**, *89*, 141.
- (54) Khandogin, J.; Brooks, C. L., III *Biochemistry* **2006**, *45*, 9363.
- (55) Kong, X.; Brooks, C. L., III *J. Chem. Phys.* **1996**, *105*, 2414.
- (56) Knight, J. L.; Brooks, C. L., III *J. Comput. Chem.* **2009**, *30*, 1692.
- (57) Guo, Z.; Brooks, C. L., III; Kong, X. *J. Phys. Chem. B* **1998**, *102*, 2032.
- (58) Wallace, J. A.; Shen, J. K. *Methods Enzymol.* **2009**, *466*, 455.
- (59) Wallace, J. A.; Wang, Y.; Shi, C.; Pastoor, K. J.; Nguyen, B. L.; Xia, K.; Shen, J. K. *Proteins* **2011**, *79*, 3364.
- (60) Wallace, J. A.; Shen, J. K. *J. Chem. Theory Comput.* **2011**, *7*, 2617.
- (61) Donnini, S.; Tegeler, F.; Groenhof, G.; Grubmüller, H. *J. Chem. Theory Comput.* **2011**, *7*, 1962.
- (62) Zheng, L.; Chen, M.; Yang, W. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 20227.
- (63) Khandogin, J.; Chen, J.; Brooks, C. L., III *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 18546.
- (64) Khandogin, J.; Brooks, C. L., III *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 16880.
- (65) Khandogin, J.; Raleigh, D. P.; Brooks, C. L., III *J. Am. Chem. Soc.* **2007**, *129*, 3056.
- (66) Dlugosz, M.; Antosiewicz, J. M. *J. Phys. Chem. B* **2005**, *109*, 13777.
- (67) Machuqueiro, M.; Baptista, A. M. *Biophys. J.* **2007**, *92*, 1836.
- (68) Campos, S. R.; Machuqueiro, M.; Baptista, A. M. *J. Phys. Chem. B* **2010**, *114*, 12692.
- (69) Shen, J. K. *Biophys. J.* **2010**, *99*, 924.
- (70) Wallace, J. A.; Shen, J. K. *Biochemistry* **2010**, *49*, 5290.
- (71) Mongan, J.; Case, D. A. *Curr. Opin. Struct. Biol.* **2005**, *15*, 157.
- (72) Knight, J. L.; Brooks, C. L., III *J. Chem. Theory Comput.* **2011**, *7*, 2728.
- (73) Knight, J. L.; Brooks, C. L., III *J. Comput. Chem.* **2011**, *32*, 3423.
- (74) Onufriev, A.; Case, D. A.; Ullmann, G. M. *Biochemistry* **2001**, *40*, 3413.
- (75) Klingen, A. R.; Bombarda, E.; Ullmann, G. M. *Photochem. Photobiol. Sci.* **2006**, *5*, 588.
- (76) Brooks, B. R.; Brooks, C. L., III; Mackerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caffisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. *J. Comput. Chem.* **2009**, *30*, 1545.
- (77) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.
- (78) Feig, M.; Karanicolas, J.; Brooks, C. L., III *J. Mol. Graphics Modell.* **2004**, *22*, 377.
- (79) Mackerell, A. D.; Denning, E. J.; Priyakumar, U. D.; Nilsson, L. *J. Comput. Chem.* **2011**, *32*, 1929.
- (80) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327.
- (81) Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 490.
- (82) Khavrutskii, I. V.; Price, D. J.; Lee, J.; Brooks, C. L., III *Protein Sci.* **2007**, *16*, 1087.
- (83) Albery, R. A.; Smith, R. M.; Bock, R. M. *J. Biol. Chem.* **1951**, *193*, 425.
- (84) Cavalieri, L. F. *J. Am. Chem. Soc.* **1953**, *75*, 5268.
- (85) Hamelberg, D.; Mongan, J.; McCammon, J. A. *J. Chem. Phys.* **2004**, *120*, 11919.
- (86) Wu, X. W.; Brooks, B. R. *Chem. Phys. Lett.* **2003**, *381*, 512.
- (87) Wu, X.; Brooks, B. R. *J. Chem. Phys.* **2011**, *134*, 134108.
- (88) Zheng, L.; Chen, M.; Yang, W. *J. Chem. Phys.* **2009**, *130*, 234105.
- (89) Lee, M. S.; Salsbury, F. R., Jr.; Olson, M. A. *J. Comput. Chem.* **2004**, *25*, 1967.
- (90) Lee, M. S.; Olson, M. A. *J. Phys. Chem. B* **2005**, *109*, 5223.
- (91) Wagoner, J. A.; Pande, V. S. *J. Chem. Phys.* **2011**, *134*, 214103.

- (92) Beglov, D.; Roux, B. *Biopolymers* **1995**, *35*, 171.
- (93) Hamaneh, M. B.; Buck, M. *J. Comput. Chem.* **2009**, *30*, 2635.
- (94) Warshel, A.; King, G. *Chem. Phys. Lett.* **1985**, *121*, 124.
- (95) King, G.; Warshel, A. *J. Chem. Phys.* **1989**, *91*, 3647.
- (96) Luzhkov, V.; Warshel, A. *J. Comput. Chem.* **1992**, *13*, 199.
- (97) Fan, Z. Z.; Hwang, J. K.; Warshel, A. *Theor. Chem. Acc.* **1999**, *103*, 77.
- (98) Chen, J.; Brooks, C. L., III; Khandogin, J. *Curr. Opin. Struct. Biol.* **2008**, *18*, 140.
- (99) Chocholousová, J.; Feig, M. *J. Phys. Chem. B* **2006**, *110*, 17240.
- (100) Liu, L.; Cottrell, J. W.; Scott, L. G.; Fedor, M. *J. Nat. Chem. Biol.* **2009**, *5*, 351.
- (101) Samijlenko, S. P.; Krechkivska, O. M.; Kosach, D. A.; Hovorun, D. M. *J. Mol. Struct.* **2004**, *708*, 97.
- (102) Harańczyk, M.; Rak, J.; Gutowski, M. *J. Phys. Chem. B* **2005**, *109*, 11495.
- (103) Bachorz, R. A.; Rak, J.; Gutowski, M. *Phys. Chem. Chem. Phys.* **2005**, *7*, 2116.

Conformational Transitions and Convergence of Absolute Binding Free Energy Calculations

Mauro Lapelosa, Emilio Gallicchio,* and Ronald M. Levy*

BioMaPS Institute for Quantitative Biology and Department of Chemistry and Chemical Biology, Rutgers the State University of New Jersey, Piscataway, New Jersey 08854, United States

ABSTRACT: The Binding energy distribution analysis method (BEDAM) is employed to compute the standard binding free energies of a series of ligands to a FK506 binding protein (FKBP12) with implicit solvation. Binding free energy estimates are in reasonably good agreement with experimental affinities. The conformations of the complexes identified by the simulations are in good agreement with crystallographic data, which were not used to restrain ligand orientations. The BEDAM method is based on λ -hopping Hamiltonian parallel replica exchange (HREM) molecular dynamics conformational sampling, the OPLS-AA/AGBNP2 effective potential, and multistate free energy estimators (MBAR). Achieving converged and accurate results depends on all of these elements of the calculation. Convergence of the binding free energy is tied to the level of convergence of binding energy distributions at critical intermediate states where bound and unbound states are at equilibrium, and where the rate of binding/unbinding conformational transitions is maximal. This finding mirrors similar observations in the context of order/disorder transitions as for example in protein folding. Insights concerning the physical mechanism of ligand binding and unbinding are obtained. Convergence for the largest FK506 ligand is achieved only after imposing strict conformational restraints, which however require accurate prior structural knowledge of the structure of the complex. The analytical AGBNP2 model is found to underestimate the magnitude of the hydrophobic driving force toward binding in these systems characterized by loosely packed protein–ligand binding interfaces. Rescoring of the binding energies using a numerical surface area model corrects this deficiency. This study illustrates the complex interplay between energy models, exploration of conformational space, and free energy estimators needed to obtain robust estimates from binding free energy calculations.

INTRODUCTION

Molecular recognition is essential for virtually all biological processes. By binding to specific sites, medicinal compounds modulate the specific activity of protein targets; the main aim of structure-based drug design is to select compounds with both specific and strong affinity for their target receptors. Accurate computational prediction of the affinity of small molecules to proteins remains a very difficult task.^{1–3} Docking programs that predict the orientation of a small molecule in the three-dimensional structure of the receptor have become a widely used tool for structure-based rational drug design.^{4–7} Docking and scoring approaches are useful to screen large databases of ligand candidates but are not considered sufficiently accurate for quantitative estimation of the binding free energies.⁸ One reason is that docking-based methods do not generally treat entropic effects and receptor flexibility, which have a significant effect on binding affinities. Physics based-models for binding,^{9,10} which use realistic representations of molecular interactions and atomic motion, have the potential to include these important effects.

Statistical mechanics provides the framework to deriving a comprehensive theory for the binding free energy of ligands to a protein.¹¹ Simplified formulations of this theory, such as the linear interaction energy (LIE)^{12–14} models, have been proposed. Other so-called end point methods such as MM/PBSA and mining minima (MM)^{15,16} include explicit or implicit approximations and simplifications. Even the most advanced free energy models available based on molecular dynamics (MD) multistate sampling and state of the art atomistic force fields are affected by inaccuracies due to limitations of potential models

and conformational sampling, as well as uncertainties regarding the relevant physicochemical states of the system (solution conditions, protonation state and tautomeric state assignments, etc.).¹⁷ Despite these challenges, MD-based free energy models remain a key topic of development given their potential to achieve sufficient realism to tackle detailed aspects of ligand optimization, and to address questions such as drug specificity and resistance.

Applications of free energy methods to pharmaceutical design have historically focused on computing the relative binding free energy between two related compounds to the same receptor.^{18–21} These models, commonly based on free energy perturbation (FEP) or thermodynamic integration free energy estimators, are most suitable for sets of very similar ligands sharing the same binding mode. There are, however, many instances where one is interested in estimating binding affinities of sets of structurally diverse molecules such as when searching for novel scaffolds or comparing binding to different mutant forms of the receptor. In recent years, double decoupling free energy methodologies that allow the computation of absolute, rather than relative, binding free energies have been reported.^{22–25} Early studies in this area have often been proofs of principle,^{26,27} and recent applied work has focused on simple model systems.^{28,29} A critical aspect of these methods is the level of conformational sampling, which needs to be capable of generating a sufficiently comprehensive representation of protein–ligand conformations, including possibly rearrangements of the receptor.^{2,22,30,31}

Received: September 28, 2011

Published: December 01, 2011

In this paper, we analyze the performance of the binding energy distribution analysis method (BEDAM), an absolute binding free energy method we recently proposed,²⁹ on the calculation of the standard binding free energies of a series of inhibitors of the FKBP receptor.³² BEDAM is based on the statistical analysis of probability distributions of the effective binding energy of the complex as the function of a thermodynamic progress variable, λ , connecting the coupled and decoupled states of the complex. The methodology is aimed at not only providing reasonable affinity estimates but also at providing physical insights concerning the driving forces for or against binding. The method makes use of parallel Hamiltonian replica exchange (HREMD) to enhance conformational sampling efficiency to search for the most effective binding mode as well as to explore multiple binding modes as a function of λ . Advanced statistical reweighting techniques^{33–35} are used to optimally merge data obtained along the binding thermodynamic path. In BEDAM, solvation effects are treated using the analytical generalized Born plus nonpolar (AGBNP) implicit solvent model,^{36,37} which is particularly suitable for binding applications.

One aim of this study is to validate the BEDAM computational protocol, previously tested on fragment binding to a model binding site,²⁹ to more complex systems closer to actual pharmaceutical design. We selected a validation system composed of a series of inhibitors of a FK506 binding protein (FKBP12). FKBP12 is a 12 kD immunophilin enzyme which catalyzes the *cis*–*trans* isomerization of peptide bonds. Inhibitors bind in the relatively shallow and solvent-exposed hydrophobic pocket.³⁸ The FKBP system is a suitable target for this purpose, as it is relatively well understood and has been studied before by double decoupling free energy methods.^{39,40} The larger size of the ligands is one obvious difference between this system and the mutant T4 lysozyme system we studied previously.²⁹ As it involves alchemically creating protein–ligand interactions at the expense of hydration interactions, the magnitude of the perturbation to the system increases with ligand size. The mixed hydrophobic and polar composition of the ligands and of the receptor binding pocket, and their partial exposure to solvent, also contribute to the added complexity of this system relative to earlier tests.

We evaluate on this system a loose restraining scheme of the kind that would be employed when structural information regarding the complexes is either lacking or uncertain. The protocol employed here does not restrain the complex to the crystallographic conformation; the ligand is free to explore all orientations and a wide variety of positions in the binding site, and both the ligand and the receptor are modeled flexibly. This choice, although a source of added complexity, reflects our interest in evaluating the ability of the BEDAM method to predict the main binding mode, together with other binding modes if present, and in studying the physical mechanisms of ligand binding and unbinding. An approach which does not rely as much on crystallographic information is conceivably better suited for cases when information about the structure of the complex is uncertain or not available, or when, such as in fragment screening,^{41–43} multiple binding poses contribute to binding. On the other hand, it has been shown that including available structural information, implemented in terms of configurational restraints, leads to better convergence of binding free energies.^{17,44} Previous binding free energy studies of the FKBP12 target,^{39,40} in particular, employed a conservative restraining

approach, using the knowledge of the structures of the complexes to restrain sampling near the bound structures of the protein–ligand complexes. In this study, we address the relationship between the level of restraining used to define the binding site volume and the ease of convergence of absolute binding free energy calculations. The ability to follow the mechanism of association is one benefit of conducting this kind of binding free energy calculations in a less restrained fashion. We illustrate examples of association events that could be relevant to the physical mechanism of binding.

The study of the role of the binding reorganization free energy effects is also facilitated by a wider exploration of conformational space. It is often the case that both binding partners undergo substantial conformational changes upon binding. The corresponding reorganization free energy is recognized as an important component of the binding free energy⁴⁵ and a key factor needed to understand ligand affinity and specificity.⁴⁶ We have previously shown that conformational reorganization upon binding plays a pivotal role in cases, such as epitope–antibody binding,^{47,48} in which there are small variations in the binding interface and most of the variations in binding affinities are due to differences in reorganization of the binding partners. Binding reorganization effects have been shown to be important in the FKBP12 system as well.^{49,50} One question we would like to answer is to what extent the λ -hopping HREMD conformational sampling algorithm in BEDAM is capable of capturing binding reorganization effects, particularly those involving internal degrees of freedom of the receptor and the ligand, which are not directly accelerated by the method.

We find that in this regime the convergence of the binding free energy is dominated by the rate of conformational transitions between bound and unbound macrostates of the complex, and that these transitions have the features of pseudo-order/disorder phase transition analogous to protein folding equilibria. The rate of conformational transitions depends mainly on the conformational sampling algorithm, and conversely, optimization of the λ schedule does not necessarily lead to an improvement of the convergence rate of the binding free energy. We believe that these are general issues applicable to many alchemical binding free energy methods.

THEORY AND METHODS

The Binding Energy Distribution Analysis Method (BEDAM). The BEDAM method²⁹ computes the binding free energy ΔF_{AB}° for the monovalent association of a receptor *A* and a ligand *B* by means of the expression

$$\begin{aligned}\Delta F_{AB}^\circ &= -kT \ln [C^\circ V_{\text{site}} \int du p_0(u) e^{-\beta u}] \\ &= -kT \ln C^\circ V_{\text{site}} + \Delta F_{AB}\end{aligned}\quad (1)$$

which follows, without approximations, from a well-established statistical mechanics theory of molecular association,¹¹ where $\beta = 1/kT$, C° is the standard concentration of ligand molecules (set to $C^\circ = 1$ M, or equivalently 1668 \AA^{-3}), V_{site} is the volume of the binding site, and $p_0(u)$ is the probability distribution of binding energies collected in an appropriate decoupled ensemble of conformations in which the ligand is confined in the binding site while the receptor and the ligand are both

interacting only with the solvent continuum and not with each other. The binding energy

$$u(\mathbf{r}_B, \mathbf{r}_A) = V(\mathbf{r}_B, \mathbf{r}_A) - V(\mathbf{r}_B) - V(\mathbf{r}_A) \quad (2)$$

is defined for each conformation $\mathbf{r} = (\mathbf{r}_B, \mathbf{r}_A)$ of the complex as the difference between the effective potential energies $V(\mathbf{r})$ of the associated and separated conformations of the complex without conformational rearrangements. In our implementation, BEDAM employs an effective potential in which the solvent is represented implicitly by means of the AGBNP2 implicit solvent model³⁷ together with the OPLS-AA^{51,52} force field for covalent and nonbonded interatomic interactions.

Equation 1 explicitly indicates that the standard binding free energy is the sum of two terms. The first term, $-kT \ln C^\circ V_{\text{site}}$, represents the entropic work to transfer the ligand from the solution environment at concentration C° into the binding site of the complex. This term depends only on the standard state and the definition of the complex macrostate. The second term, ΔF_{AB} , involving the Boltzmann-weighted integral of $p_0(u)$, corresponds to the work for turning on the interactions between the receptor and the ligand when the ligand is confined in the binding site region.²⁹ Equation 1 also naturally leads to the definition of a binding affinity density function $k(u) = C^\circ V_{\text{site}} p_0(u) \exp(-\beta u)$ in terms of which the binding constant is written as

$$K_{AB} = e^{-\beta \Delta F_{AB}^\circ} = \int k(u) du \quad (3)$$

On the basis of eq 3, the binding affinity density $k(u)$ can be interpreted as a measure of the contribution of the conformations of the complex with the binding energy, u , to the binding constant.²⁹ We have shown that $k(u)$ is proportional to $p_1(u)$, the binding affinity density in the coupled ensemble of the complex, with a proportionality constant related to the binding free energy.²⁹

The larger the value of the integral in eq 1, the more favorable is the binding free energy. The magnitude of the $p_0(u)$ distribution at positive, unfavorable values of the binding energy u reflects the entropic thermodynamic driving force which opposes binding, whereas the tail at negative, favorable binding energies measures the energetic gain for binding due to the formation of ligand–receptor interactions. The interplay between these two opposing forces ultimately determines the strength of binding. We found that the ability of BEDAM to explicitly include both favorable energetic gains and unfavorable entropic losses to be essential to properly reproducing experimental binding affinities in a challenging set of candidate ligands to T4 lysozyme receptors whose estimates of binding affinity failed by simplified docking and scoring approaches.⁵³

The accurate calculation of the important low energy tail of $p_0(u)$ can not be accomplished by a brute-force collection of binding energy values from a simulation of the complex in the decoupled state because these are rarely sampled when the ligand is not guided by the interactions with the receptor. Instead, we use biased sampling and parallel Hamiltonian replica exchange (HREM), in which swarms of coupled replicas of the system, differing in the value of an interaction parameter $0 \leq \lambda \leq 1$ controlling the strength of ligand–receptor interactions, are simulated simultaneously. The replicas collectively sample a wide range of unfavorable, intermediate, and favorable binding energies which are then unbiased and combined together by means of reweighting techniques.^{34,35}

BEDAM is based on biasing potentials of the form $\lambda u(\mathbf{r})$, yielding a family of λ -dependent hybrid potentials of the form

$$V_\lambda(\mathbf{r}) = V_0(\mathbf{r}) + \lambda u(\mathbf{r}) \quad (4)$$

where

$$V_0(\mathbf{r}) = V(\mathbf{r}_B) + V(\mathbf{r}_A) \quad (5)$$

is the potential energy function of the decoupled state. It is easy to see from eqs 2, 4, and 5 that $V_{\lambda=1}$ corresponds to the effective potential energy of the coupled complex and $V_{\lambda=0}$ corresponds to the state in which the receptor and ligand are not interacting (decoupled state). Intermediate values of λ trace an alchemical thermodynamic path connecting these two states. The binding free energy ΔF_{AB} is by definition the difference in free energy between these two states.

For later use, we introduce here the reorganization free energy for binding $\Delta G_{\text{reorg}}^\circ$ defined by the expression¹⁰

$$\Delta F_{AB}^\circ = \langle u \rangle_1 + \Delta G_{\text{reorg}}^\circ \quad (6)$$

where $\langle u \rangle_1$ is the average binding energy at $\lambda = 1$ and ΔF_{AB}° is the standard binding free energy.

Soft Core Implementation. To improve convergence of the free energy near $\lambda = 0$, in this work, we employ a modified “soft-core” binding energy function,¹⁰ similar in spirit to a recently proposed approach,³⁴ of the form

$$u'(\mathbf{r}) = \begin{cases} u_{\text{max}} \tanh[u(\mathbf{r})/u_{\text{max}}], & u(\mathbf{r}) > 0 \\ u(\mathbf{r}) & u(\mathbf{r}) \leq 0 \end{cases} \quad (7)$$

where u_{max} is some large positive value (set in this work as 1000 kcal/mol). This modified binding energy function, which is used in place of the actual binding energy function [eq 2] wherever it appears, caps the maximum value of the binding energy while leaving unchanged the value of favorable binding energies. This serves two purposes. One purpose is to improve sampling at small λ 's by letting atoms pass through each other without clashing. This is possible because for values of λ around $1/u_{\text{max}}$ or smaller, $\lambda u'(\mathbf{r})$ in eq 4 is guaranteed to be comparable to thermal energy even for conformations with atomic overlaps and large and unfavorable binding energies. In contrast, with the original definition of the binding energy, only at $\lambda = 0$ can atoms pass through each other without clashing. In addition, the soft-core binding energy function simplifies the choice of the λ schedule near $\lambda = 0$. As discussed below, overlaps between neighboring binding energy distributions are necessary for free energy estimation. The large range of binding energies sampled in the positive range and the rate of change of the binding energy distributions as a function of λ for small λ make it difficult to select a λ schedule near $\lambda = 0$ to ensure sufficient overlaps between binding energy distributions. The distributions $p_\lambda(u')$ of the soft core binding energy are instead much better behaved because the range of soft core binding energies has a finite upper limit ($u' = u_{\text{max}}$) for any conformation of the complex. Moreover, because the corresponding replicas can sample equally well conformations with atomic overlaps, it is guaranteed that distributions $p_\lambda(u')$ for values of λ on the order of $1/u_{\text{max}}$ or smaller will overlap with the distribution at $\lambda = 0$. In this work, we set the minimum nonzero value of λ as $1/u_{\text{max}}$ and confirmed numerically that the corresponding binding energy distribution overlaps significantly with $p_0(u)$.

It is evident from eq 4 that $\lambda = 0$ identifies the decoupled state regardless of the definition of the binding energy. However, the potential energy function of the coupled state $V_1 = V_0 + u$ is in principle affected by whether we employ eq 2 or eq 7 to represent the binding energy. Therefore, we must consider to what degree the binding free energy, which measures the free energy difference between these two states, is affected by the introduction of the soft core binding energy function. Intuitively, the binding free energy cannot significantly be affected by the soft core function if u_{\max} is very large compared to thermal energy (we set $u_{\max} = 1000$ kcal/mol). The alternative would imply that the binding affinity depends on the details of the interatomic potentials at high energies which are not known accurately or, much less, modeled correctly by classical force fields. Indeed, the $\lambda = 1$ ensemble with the soft core function is virtually indistinguishable from the original $\lambda = 1$ ensemble because the probability of sampling large positive binding energies, which are the only cases in which the original and soft core binding energy functions differ substantially, is infinitesimally small at $\lambda = 1$. A study including theoretical considerations and numerical tests of hard-core versus soft-core functions will be reported in a separate publication.

Free Energy Estimation. In this work, we employ the multi-state Bennett acceptance ratio estimator (MBAR)^{35,55} to estimate binding energy distributions and standard binding free energies from binding energy samples obtained from the HREM simulations. On the basis of eq 4, the binding free energy ΔF_{AB} , which is the standard binding free energy minus that standard state term $-kT \ln C^\circ V_{\text{site}}$, is given by the free energy difference between the $\lambda = 1$ and $\lambda = 0$ states with potential energy functions defined by eq 4. This is a consequence of having selected biasing potentials, aimed at properly sampling the $p_0(u)$ distribution at low binding energies, of the form λu —noting, however, that in general BEDAM computes the binding free energy based on eq 1 where $p_0(u)$ is estimated by the application of any suitable series of biasing potentials not necessarily connecting the decoupled and coupled states. With the present setup it is nevertheless convenient to compute the binding free energy directly from the MBAR dimensionless free energies \hat{f}_λ using the relationship

$$\Delta F_{AB} = kT(\hat{f}_1 - \hat{f}_0) \quad (8)$$

The MBAR dimensionless free energies $\hat{f}_\lambda = -\ln Z_\lambda$ are defined as the negative of the logarithm of the λ -dependent biased partition functions Z_λ . In this case, the dimensionless free energies are estimated by the self-consistent solution of the set of equations³⁵

$$\hat{f}_i = -\ln \sum_{j=1}^K \sum_{n=1}^{N_j} \frac{\exp[-\beta \lambda_i u_{jn}]}{\sum_{k=1}^K N_k \exp[\hat{f}_k - \beta \lambda_k u_{jn}]} \quad (9)$$

where $\hat{f}_i = \hat{f}_{\lambda_i}$, u_{jn} is the n th binding energy sample from replica j sampled with biasing potential λ_j , K is the number of replicas, and N_j is the total number of binding energy samples from replica j . For the MBAR analysis, we employed the code provided by John Chodera and Michael Shirts.³⁵ Block bootstrap analysis⁵⁶ with 10 blocks and eight resampling trials was used to estimate statistical uncertainties.

Computational Details. BEDAM calculations were performed for seven ligands of FKBP (Figure 1). This is the same ligand series investigated in prior binding free energy studies.^{39,40}

The initial structure of the complexes with ligands 8, 9, and 20 were retrieved from the PDB (PDB IDs 1FKG, 1FKH, and 1FKJ, respectively). The starting structures for the receptor of the other complexes were built on the basis of the existing crystal structure of 1FKG. The crystallographic water molecules and ions were removed. Ionization states were assigned considering neutral pH. Histidines were not protonated, and hydrogen atoms were added. $C\alpha$ atoms of the FKBP receptor were restrained around the X-ray original coordinates using an isotropic quadratic function with force constant $k_f = 0.6$ kcal/mol/Å², which allows a motion of 4 Å around the original positions at the temperature used in the simulations. Free motion of all of the remaining atoms was allowed. This restraining scheme leads to convergence of the binding free energies while it is sufficiently relaxed to encompass all likely conformations of the complex. All of the simulations were carried out using the IMPACT program.⁵⁷ The RMSDs of the ligand conformations were calculated including the atoms of the core (C1–C2–C3–C4–C5–C6–N7–C8) only.

All of the complexes were minimized and thermalized at 300 K using the same procedure. λ -biased molecular dynamics simulations were performed for 2 ns per replica (~280 ns of total computation time for all complexes) with a time step of 1.5 fs at 300 K. Fifteen replicas at $\lambda = 0, 10^{-3}, 2 \times 10^{-3}, 4 \times 10^{-3}, 6 \times 10^{-3}, 8 \times 10^{-3}, 10^{-2}, 2 \times 10^{-2}, 6 \times 10^{-2}, 0.1, 0.25, 0.5, 0.75, 0.9$, and 1 were used for all ligands. For ligand 20, we also employed 36 replicas at $\lambda = 0, 10^{-3}, 2 \times 10^{-3}, 4 \times 10^{-3}, 6 \times 10^{-3}, 8 \times 10^{-3}, 10^{-2}, 2 \times 10^{-2}, 6 \times 10^{-2}, 0.1, 0.25, 0.30, 0.45, 0.50, 0.55, 0.60, 0.62, 0.65, 0.68, 0.70, 0.72, 0.74, 0.75, 0.77, 0.80, 0.82, 0.85, 0.86, 0.87, 0.88, 0.89, 0.90, 0.92, 0.94, 0.97$, and 1.0. We employed the OPLS-AA⁵¹ force field with the AGBNP2 implicit solvent model.^{36,37} The AGBNP2 model includes a novel first solvation shell hydration function to improve the balance between solute–solute and solute–solvent interactions that makes it more suitable for free energy calculations. Bond lengths with hydrogen atoms were constrained using SHAKE. A 12 Å residue-based cutoff was imposed on both direct and generalized Born pair interactions. Binding energies for protein–ligand binding were calculated every 1 ps during the second nanosecond of the simulation. The data set for each complex consisted of 15 000 binding energy values corresponding to 1000 samples for each of 15 HREMD replicas.

The replicas were coupled using a Hamiltonian replica exchange method (HREMD) previously described, which was shown to significantly improve conformational sampling efficiency.²⁹ As further discussed below, the λ schedule was determined so as to ensure overlaps between neighboring binding energy distributions and frequent accepted λ exchanges (acceptance ratio of at least 50%).

The binding site volume^{11,58} was defined in terms of flat-bottom harmonic potentials as previously described.²⁹ Briefly, an indicator function is introduced of the form $I(r, \cos \theta, \phi) = \exp[-\beta \omega(r, \cos \theta, \phi)]$ where $\omega(r, \cos \theta, \phi)$ is a product of flat-harmonic potentials acting on the position, expressed in polar coordinates, of a reference atom of the ligand with respect to the position of the $C\alpha$ atoms of three reference residues of the receptor.⁵⁹ The distance restraint potential between atom C2 of the ligand (see Figure 1) and the $C\alpha$ atom of residue 55 was centered at a 5 Å distance with a 5 Å tolerance on either side; beyond these limits a quadratic function penalizes the distances with a force constant of 3 kcal/mol/Å². The flat-bottom harmonic restraint potential for the cosine of the angle θ between the reference ligand atom, the $C\alpha$ atom of residue 55, and the $C\alpha$

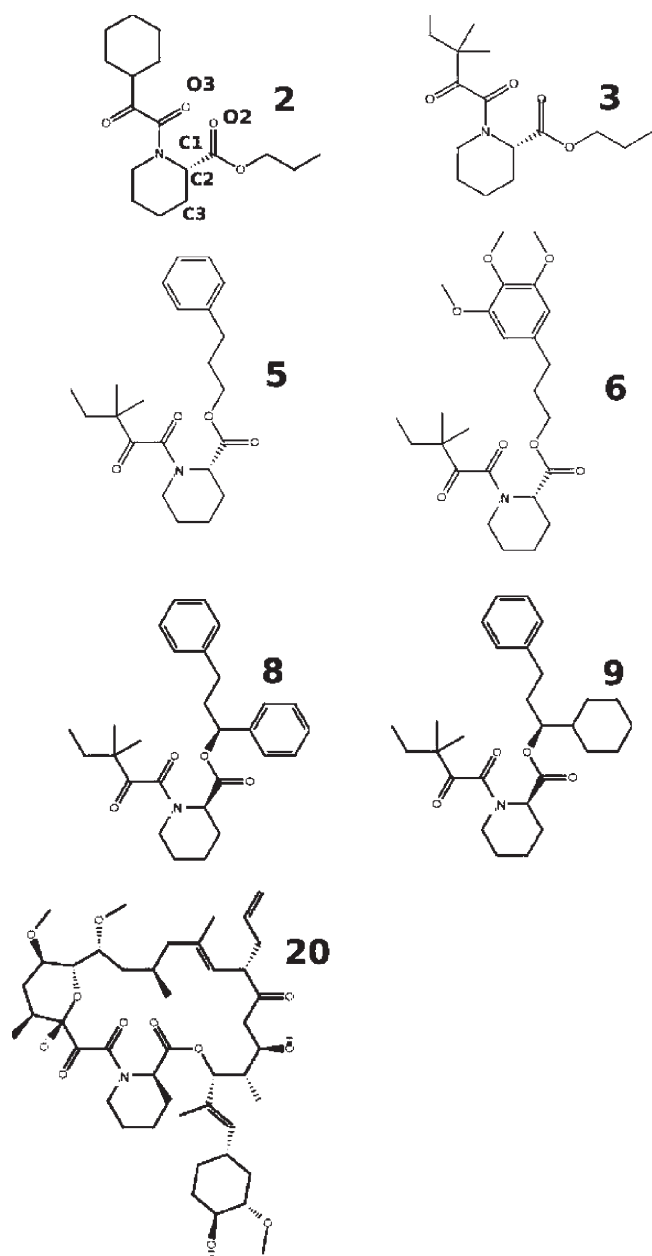


Figure 1. Structural formulas of the seven ligands of FKBP12 investigated in this work. Ligand 20 (bottom right) also known as Tacrolimus or FK506 is an immunosuppressive drug. Atom labels are shown for ligand 2.

atom of residue 46 was centered at $\cos \theta = 0.67$ with a 0.3 tolerance on both sides and a force constant of 100 kcal/mol beyond that. The restraint potential for the dihedral angle defined by the three atoms described above plus the $C\alpha$ atom of residue 44 was centered at $\phi = 32^\circ$ with a tolerance of 50° on either side and a force constant of 0.1 kcal/mol/deg beyond that. The volume of the binding site ($V_{\text{site}} = 504 \text{ \AA}^3$) given by the integral of the indicator function as defined was computed analytically. The resulting standard state term is computed as $-kT \ln C^\circ V_{\text{site}} = 0.71 \text{ kcal/mol}$. For ligand 20, we also employed a stricter definition of the complex by imposing additional limits on the orientation of the ligand in the binding site. These were implemented in terms of flat-bottom harmonic

potentials similar to the those given above based on one bond angle and two dihedral angles involving two additional reference atoms (C1 and C3, see Figure 1) of the ligand as described⁵⁹ centered on the crystallographic orientation (1FKJ). The volume of this more restrictive binding site volume and the corresponding standard state free energy term were computed as 43 \AA^3 and 2.71 kcal/mol , respectively.

A rescoring procedure was conducted to overcome a deficiency of the AGBNP2 surface area model. In AGBNP2, the free energy of cavity formation is modeled in terms of the solute surface area, which is calculated taking the analytical derivative of the solute volume. This analytical model is fast and yields stable MD trajectories but is not particularly accurate for these systems relative to numerical molecular surface or solvent accessible surface area evaluations.³⁷ As sufficiently accurate surface area models with the desired characteristics are currently lacking, in this work, we sought to replace in postprocessing the cavity term, denoted here as $G_{\text{cav}}(1)$, of the AGBNP2 model³⁷ with a more accurate model $G_{\text{cav}}(2) = \gamma_2 A_{\text{SASA}}$ where A_{SASA} is the solvent-accessible surface area of the solute evaluated numerically (we used the SURFV⁶⁰ program) and $\gamma_{(2)}$ is an adjustable surface tension parameter. We estimated the change in binding free energy, $\Delta\Delta F_{AB}$, on going from the original cavity free energy model to the numerical SASA cavity free energy model assuming first order perturbation theory, by rescoring the binding energies of the conformations of the complex collected at $\lambda = 1$:

$$\Delta\Delta F_{AB} \cong \langle u_{(2)} - u_{(1)} \rangle_1 = \gamma_{(2)} \langle \Delta A_{\text{SASA}} \rangle_1 - \langle \Delta G_{\text{cav}}(1) \rangle_1 \quad (10)$$

where $u_{(1)}$ and $u_{(2)}$ represent, respectively, the binding energies of each complex conformation evaluated with the original and numerical cavity free energy models, $\Delta A_{\text{SASA}} = A_{\text{SASA}}(AB) - A_{\text{SASA}}(A) - A_{\text{SASA}}(B)$ is the change in surface area on going from the separated ligand and receptor to the complexed conformation, and $\Delta G_{\text{cav}}(1)$ is the corresponding change in cavity free energy as computed using the original model. To assign a value for the surface tension coefficient $\gamma_{(2)}$, eq 10 was fitted to the residuals between the binding free energies computed with the original energy model and the experimental affinities. We obtained $\gamma_{(2)} = 0.051 \text{ kcal/mol/\AA}^2$, a value in good agreement with an earlier independent analysis of cavity hydration free energies.⁶¹

RESULTS

FKBP Ligands. This study covers a ligand set composed of seven related inhibitors (Figure 1) of FKBP12 (Figure 2) with known experimental binding free energies.³² These ligands were originally developed by stepwise addition of hydrophobic rings to a central core in an attempt to increase potency. The compounds contain from one to three rings and several different functional groups commonly encountered in drug-like molecules. The inhibitors have moderate molecular weight (~ 450 – 500 u), with the exception of FK506 (ligand 20 in Figure 1), a natural inhibitor of FKBP12 with an unusually large molecular weight (804 u) compared to most drug-like molecules. The measured binding free energies of the compounds range from low nanomolar to micromolar (Table 1). Crystal structures are available for three of the inhibitors (1FKG for ligand 8, 1FKH for ligand 9, and 1FKJ for ligand 20).³² These show close similarity between the

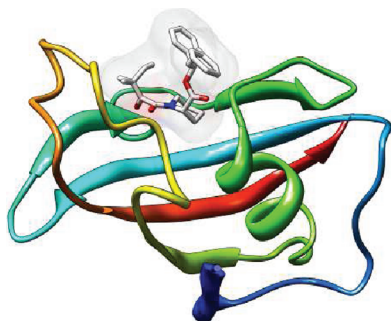


Figure 2. Representation of the crystal structure (PDB ID 1FKJ) of the complex of FKBP12 with ligand 9. The receptor is shown in cartoon representation and the ligand in wire representation.

binding modes of the core of the ligands, as represented in Figure 2 for ligand 9. Two carbonyl oxygen atoms, O2 and O3 (see Figure 1, ligand 2, for atom labeling), of the ligand core form two hydrogen bonds with Tyr82 and Ile56. Hydrophobic moieties form π - π contacts with Tyr82 and Phe46, and other generally hydrophobic contacts with Gln53, Glu54, Val55, Ile56, His87, and Ile90 of the receptor.

Binding Free Energy Estimates. The computed binding free energies obtained from BEDAM (see Methods) are shown in the second column of Table 1 compared to the available measurements.³² As further discussed below, because convergence could not be achieved with the standard settings, the calculated binding free energy for ligand 20 reported in Table 1 was obtained using a different, stricter, binding site definition than for the other ligands. The computed affinities achieve good discrimination between good binders and weak binders. The confidence ranges reported in Table 1, estimated using the bootstrap method, indicate robustness of the free energy estimates with respect to variations in the binding energy data. However, these probably underestimate the actual statistical uncertainties due to the difficulty of reliably inferring them from a finite set of samples.

The calculated standard binding free energies are reasonably correlated with the experimental values ($R^2 = 0.59$, not including ligand 20). However, it is clear that the magnitude of the binding free energies is underestimated by the BEDAM model. A likely cause of the discrepancy is the inaccurate representation of changes in nonpolar solvation upon binding. To test this, we have computed the surface area loss upon binding using an accurate numerical method (see Methods) for all of the trajectory frames corresponding to the coupled state ($\lambda = 1$) and compared these to the surface area loss estimates produced by AGBNP2. We found that on average the AGBNP2 surface area model captures only about half of the surface area loss. This result is consistent with the nature of the analytical surface area model implemented in AGBNP2, which is based on an atomic radius offset of 0.5 Å, significantly smaller than the conventional value of 1.4 Å commonly used for the solvent accessible surface. Because the surfaces of each of the binding partners do not sufficiently extend over the region between the two molecules, the AGBNP2 analytical surface model incorrectly represents as solvent exposed some of the atoms facing voids within the binding interface too small to contain water molecules. This results in the underestimation of the surface area loss of these atoms as they go from solvent-exposed in the conformation to buried in the bound conformation. Using the procedure described in the Methods section, we have rescored the binding free

Table 1. Experimental and Calculated Standard Binding Free Energies of the Complexes of FKBP12 with the Ligands Investigated^a

ligand	exptl ^b	calcd ^b	S-Calc ^b	$\langle u \rangle_1^b$	$\Delta G_{\text{reorg}}^c$
2	-7.80 ± 0.1	-2.54 ± 0.05	-7.61 ± 0.05	-28.14	20.53
3	-8.40 ± 0.1	-3.97 ± 0.03	-7.83 ± 0.04	-28.40	20.57
5	-9.50 ± 0.1	-3.95 ± 0.04	-7.91 ± 0.04	-30.44	22.53
6	-10.80 ± 0.3	-3.82 ± 0.05	-10.95 ± 0.05	-34.47	23.52
8	-10.90 ± 0.1	-4.80 ± 0.01	-10.66 ± 0.02	-35.97	25.31
9	-11.10 ± 0.2	-6.25 ± 0.02	-12.55 ± 0.03	-36.05	23.50
20 ^c	-12.70 ± 0.2	-0.93 ± 0.05	-13.43 ± 0.03	-42.05	28.62

^aThe “S-Calc” column lists the surface area-corrected binding free energies (see text). The last two columns report the decomposition of the S-Calc estimates into the average binding energy and reorganization free energy, respectively. ^bIn kcal/mol. ^cObtained with a stricter binding site definition (see text).

energies using a more accurate surface area model and determined an optimal value of $\gamma = 0.051$ kcal/mol/Å² for the surface tension coefficient. This value is in good agreement with the surface tension coefficient to $\gamma = 0.058$ kcal/mol/Å² obtained from the cavity hydration free energies of a series of alkanes.⁶¹ This observation further supports the hypothesis that the discrepancies between the measured and computed affinities of the FKBP inhibitors is physically related to the inaccurate representation of hydrophobic driving forces, and that a more accurate geometrical description of the surface area loss upon binding leads to better quantitative predictions.

Indeed, the surface-rescored binding free energies, shown in Table 1, are in overall good agreement with the experimentally measured affinities, and the correlation with the experiments is high ($R^2 = 0.88$) as shown in Figure 3. Very good agreement is obtained for ligand 2 with a computed binding free energy of -7.84 kcal/mol compared to the experimental value of -7.80 kcal/mol (Table 1). This ligand resulted as an outlier in earlier work.³⁹ The binding free energies of ligands 3 and 4 follow the same trend toward stronger affinities as in the experiments but with significantly smaller variations. Indeed, ligands 2, 3, and 5 are estimated as having nearly equivalent affinities compared to an experimental spread of approximately 1.7 kcal/mol. The higher affinities of the second set of ligands (6, 8, and 9) are well reproduced, with the exception of the strongest binder, ligand 9, whose computational estimate (-12.7 kcal/mol) is off by 1.6 kcal/mol relative to the experimental binding free energy (-11.1 kcal/mol). The agreement for ligand 20 is very good; a large surface area correction is obtained for this ligand, consistent with its much larger surface area relative to the other ligands. The trend observed experimentally toward higher affinities in relation to an increase in size and number of cycles on the ester functionality of the inhibitors is reproduced by the calculations. The predicted rank order is in very good agreement with the measurements, with the exception of the inversion of ligands 6 and 8, which have very similar predicted binding free energies, but the experimental values differ by ~ 1 kcal/mol. Overall, the quality of the agreement between calculations and experimental measurements is sufficiently good to give us confidence that the computational model appropriately includes the main physical driving forces and that it can be used

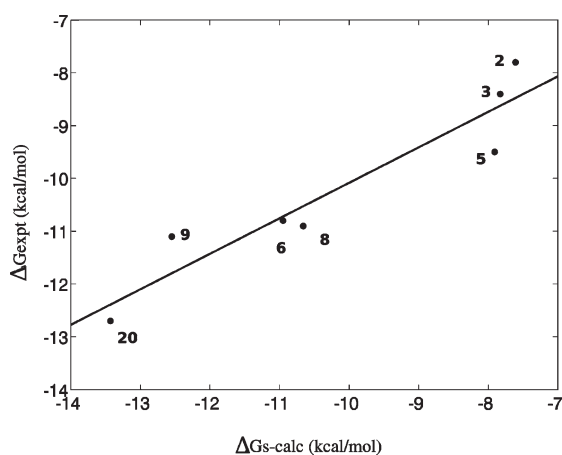


Figure 3. Correlation plot between calculated and experimental standard binding free energies of the seven FKBP complexes studied. The line represents a least-squared fit to the data.

to extract useful insights into the thermodynamics and mechanism of the binding process.

Thermodynamic Decomposition. Thermodynamic decomposition¹⁰ (see Methods) of the computed binding free energies into the average binding energy in the coupled state ($\langle u \rangle_1$) and the reorganization free energy ($\Delta G_{\text{reorg}}^\circ$; Table 1) reveals that the trend toward greater affinities on going from ligand 2 to ligand 20 is primarily driven by stronger ligand–receptor interactions. For example, the computed average binding energy $\langle u \rangle_1$ of ligand 2 is approximately -28 kcal/mol compared to -42 kcal/mol for ligand 20. This trend is consistent with the increased number of mainly hydrophobic contacts between the ester side chain of the larger ligands and the receptor.

This favorable energetic driving force toward binding is opposed by a reorganization free energy loss (Table 1), which increases in magnitude with increasing ligand size. The reorganization free energy term measures the unfavorable work necessary to remodel the unbound conformational ensembles of the ligand and the receptor in order to form favorable interactions in the bound state, and it necessarily opposes binding because in the absence of receptor–ligand interactions the ligand and the receptor would spontaneously return to their unbound states at lower free energy. The reorganization free energy cost can in turn be thought of as originating from both configurational entropy losses and energetic strain of the ligand and receptor in solution, while thermodynamic effects due to the solvent are implicitly included by the implicit free energy model of solvation.^{10,62} In the context of an implicit solvent model, the entropic cost is in part due to the loss of translational and orientational freedom of the ligand as it is being localized into the binding site of the receptor. The receptor and especially the larger and more flexible ligands undergo additional conformational entropy losses to mutually adapt their conformations in order to form favorable interactions. Bound conformations can also be disfavored energetically relative to their more relaxed unbound conformations in solution. This energetic strain opposes binding and further reduces the effect of favorable receptor–ligand interactions. As the data in Table 1 show, the reorganization free energy grows nearly monotonically from ligand 2 to ligand 20, consistent with increasing ligand size and flexibility.

Choice of λ Schedule. The choice of the number of replicas and their λ assignments affects BEDAM calculations in two

related ways. To estimate the binding free energy, it is believed to be necessary that an unbroken sequence of overlaps between the binding energy distributions $p_\lambda(u)$ be constructed between $\lambda = 0$ (the decoupled state) and $\lambda = 1$ (the coupled state). So the choice of the λ schedule must meet this minimum requirement.⁶³ A larger density of replicas however is beneficial because it increases the amount of overlap between the distributions and allows us to obtain reliable free energy estimates with fewer samples, especially when using multistate approaches such as MBAR.³⁵ The choice of the λ schedule also affects the acceptance ratio of λ exchanges in the HREM conformational sampling scheme. λ exchanges are accepted with probability $\min[1, \exp(-\beta\Delta\lambda\Delta u)]$ ²⁹ where $\Delta\lambda$ is the difference in λ 's being exchanged and Δu is the difference in binding energies between the replicas exchanging them. Statistically, the magnitude of both of these quantities tends to decrease as overlaps between binding energy distributions increase, thereby making exchanges more likely. It follows that monitoring the extent of diffusion in λ space of HREM replicas is also equivalent to monitoring the level of overlaps between binding energy distributions and ultimately the quality of the selected λ schedule.

Analysis of the HREM data shows that λ space is well explored by most of the 15 replicas of ligands 2–9, although exchange bottlenecks can be seen at specific λ 's as for example with ligand 2 between $\lambda = 0.5$ and 0.75 , consistent with the relatively small overlap between the corresponding binding energy distributions (Figure 4). In contrast, the diffusion in λ space of the complex with ligand 20 is very limited. As shown in Figure 5A, replicas very rarely cross the large gap in λ space between $\lambda = 0.5$ and $\lambda = 0.75$. The reason for this is that the corresponding distributions do not overlap to any significant extent. This is expected because the complex with ligand 20 explores a wider range of binding energies requiring more intermediate λ 's to cover it appropriately. We found that 36 replicas were sufficient to remove gaps in λ exchanges (Figure 5B) for this complex. Nevertheless, it is clear from the patterns of color mixing in Figure 5B that, unlike those for ligands 2–9, replicas for ligand 20 are divided into two disjoint groups: those that tend to visit only lower values of λ and those that tend to visit only high values of λ . This is because, even though local λ exchanges are promoted by a larger distribution of overlaps, global diffusion of replicas in λ space also depends on the ability of replicas to undergo conformational transitions.

The binding site indicator function as defined (see Methods) does not restrict configurational and rotational degrees of freedom of the ligand, and the binding site volume is sufficiently wide for the occurrence of conformations of the complex distinct from the bound crystallographic binding mode, as for example shown in Figure 7 in terms of RMS deviation. Examples of these conformations, which we refer to as unbound, are shown in Figure 6A and B. (The bound/unbound macrostates of the complex should not be confused with the coupled, $\lambda = 1$, and uncoupled, $\lambda = 0$, thermodynamic states of the complex.) Replicas in unbound conformations with unfavorable binding energies tend to remain at small values of λ , whereas replicas in bound conformations with favorable binding energies tend to remain at large values of λ . To explore the whole range of λ 's, a replica needs to undergo conformational transitions from bound to unbound conformations or vice versa.⁶⁴ This issue, which strongly affects the convergence of binding free energy estimates, is further discussed below. We note here that, because it is mainly tied to the occurrence of conformational transitions, this type of

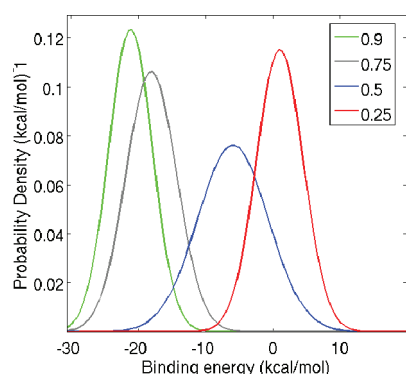


Figure 4. Binding energy probability densities for the complex with ligand 2 at $\lambda = 0.25, 0.5, 0.75,$ and 0.9 . The amount of overlap between these distributions, which is important for accurate binding free energy estimation, is reasonably good with the distribution at the critical value $\lambda_{1/2} = 0.5$ being wider than the others and acting as a bridge between the other distributions.

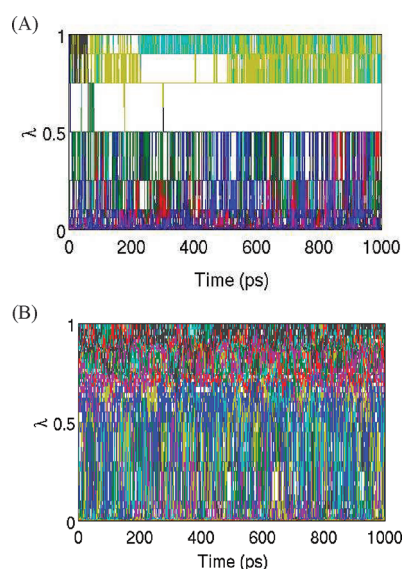


Figure 5. Time evolution of λ for each of the HREM replicas of ligand 20 with 15 replicas (A) and with 36 replicas (B). Each color corresponds to a different replica.

convergence behavior is not directly addressable by simply increasing the density of λ replicas. It must be concluded therefore that an appropriate choice of the λ schedule is a necessary but not sufficient condition for obtaining converged binding free energy estimates.

DISCUSSION

The complexes of FKBP12 we investigated in this work provide very useful data to better understand the features and behavior of the BEDAM computational protocol and alchemical binding free energy calculations in general. One of the aims has been to explore the application of the method to pharmaceutical targets involving larger and more flexible ligands than the T4 lysozyme system we have previously studied.²⁹ In this respect, the FKBP12 system is relatively well understood and has served as a useful validation target for absolute binding free energy

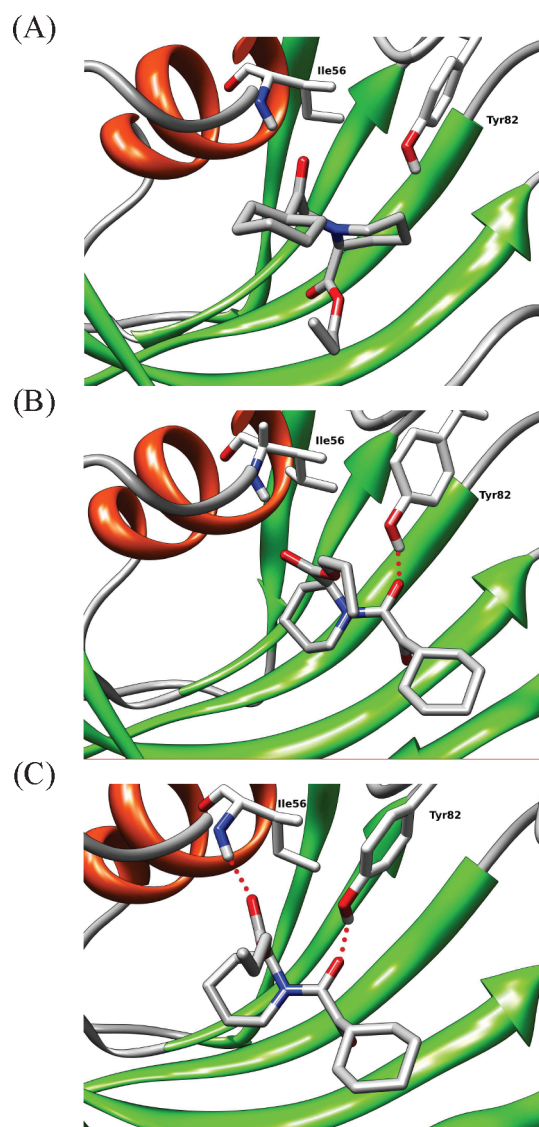


Figure 6. Transition mechanism from unbound conformations with no hydrogen bond (A) to the bound conformation with two hydrogen bonds in ligand 2 (C). The intermediate state has one hydrogen bond (B).

methods.^{39,40} The size and diversity of this ligand series (Figure 1) makes it unsuitable for relative binding free energy methods, which are more commonly employed than absolute binding free energy methods in applied research.¹ Although they are recognized as being more challenging, absolute binding free energy models are considered more suitable to study the fundamental thermodynamic components of the binding equilibrium.³ In particular, BEDAM emphasizes effects such as conformational entropy and reorganization and the contribution of multiple binding modes,²⁹ issues that are not easily tackled with methods based on relative free energy perturbation approaches.

As is often recognized, the performance of atomistic computational models primarily hinges on the quality of the energy function and the extent of conformational sampling. We have confirmed a weakness of the AGBNP2 surface-based cavity free energy model, which underestimates the favorable hydrophobic component of binding. This limitation, which also affected our previous binding free energy estimates for the T4 lysozyme

system,²⁹ is more noticeable in the present application given the larger amount of buried surface area involved. We were able to show that in this case a simple rescoring scheme of the binding energies in the coupled ensemble with a more accurate surface model is sufficient to recover binding free energies in good agreement with the experimental affinities. The validity of the surface area rescoring approach is further supported by the fact that it leads to an estimated value of the effective surface tension coefficient in agreement with independent estimates of the same quantity based on computed cavity free energies of alkanes.⁶¹ The general applicability of the surface rescoring scheme we employed here is uncertain. The method is implicitly based on first order perturbation theory, which assumes that changes in the surface area model affect binding energies without significantly altering conformational ensembles. It is conceivable therefore that these assumptions are not as valid for marginally stable complexes, or complexes characterized by multiple binding poses, whose conformational distributions are easily perturbed by changes in the potential energy model.

The BEDAM method employs advanced conformational sampling based on Hamiltonian replica exchange (HREM), a well established strategy to improve sampling in a variety of applications,^{65–67} including binding free energy calculations.^{2,68} We have shown²⁹ that HREM λ -hopping is in general quite efficient at exploring intermolecular degrees of freedom and specifically the position and orientation of the ligand relative to the receptor. This is understandable since the λ perturbation parameter directly controls the magnitude of the interaction between the ligand and the receptor. At $\lambda \approx 0$, where protein–ligand interactions are very weak, the ligand is free to explore a wide variety of positions and orientations, some of which, by means of λ exchanges, anneal to low energy conformations at $\lambda \approx 1$. Conversely, stable conformations of the complex that would normally remain trapped using conventional molecular dynamics have the opportunity to escape by migrating to smaller λ values. Overall, we have confirmed on this system the critical advantages afforded by the λ -hopping strategy at aiding conformational transitions and speeding up the convergence of free energy estimates. On the other hand, we found that binding/unbinding conformational transitions were hampered by slow conformational rearrangements not directly accelerated by λ hopping. As discussed in detail below, we found that this effect slows convergence, and it turned out to be sufficiently severe for FKS06 (ligand 20) to prevent reaching convergence for this ligand using the same protocol used for the smaller ligands.

Binding/Unbinding Transitions. As discussed,²⁹ BEDAM binding free energies rely on the probability distributions, $p_\lambda(u)$, of the binding energy as a function of λ (see Figure 4), and consequently, convergence of binding free energies is necessarily tied to the level of convergence of binding energy distributions. In principle, all of the distributions along λ are required to reach a sufficient level of convergence to achieve convergence of the binding free energy. In practice, we found that it is more difficult to converge a $p_\lambda(u)$ distribution which contains components from both bound and unbound conformations. In these cases, it is necessary to sample both macrostates with the correct probability in order to reach convergence. So, in other words, convergence of binding energy distributions depends on the quality of sampling along conformational degrees of freedom that can be considered orthogonal to the progress parameter λ .^{64,69}

We have found little evidence of multiple binding modes at $\lambda = 1$ for the FKBP12 complexes we have investigated in this

study. The conformations of the coupled state we have obtained are all characterized by the dual hydrogen bonding pattern seen in the available crystal structures as discussed above. Furthermore, we observed little variation of the distributions near $\lambda = 1$ as a function of simulation length. We conclude therefore that the distributions at λ near 1 correspond to a single, well sampled conformational macrostate and are appropriately converged. The distributions at the other end of the spectrum near $\lambda = 0$ are similarly converged. Obviously, these result from a large variety of conformations which, however, are rapidly interconverting due to weak protein–ligand interactions as $\lambda \rightarrow 0$. We find that slow convergence is instead caused by sluggish binding/unbinding conformational interconversions at intermediate critical values of λ . At these λ states, conformations of the ligand bound to the receptor in the crystallographic binding mode are in equilibrium with unbound conformations in which the ligand is either displaced from the binding site or oriented so that it is unable to form the proper interactions with the receptor. See Figure 6 for representative bound and unbound conformations.

An illustration of the conformational landscape characterizing the binding/unbinding equilibrium is presented in Figure 7 for ligand 2 at $\lambda = 0.5$. Points in the upper right of the plot with high RMSD and less favorable binding energies correspond to unbound conformations while the tight cluster of points at low RMSD and more favorable binding energies correspond to bound conformations. At this particular λ , the populations of the bound and unbound conformational macrostates are approximately equal (see Figure 8). As Figure 7A illustrates, the unbound macrostate is characterized by a wider variety of conformations spanning many units of RMSD from the reference crystallographic conformation. The interpretation is therefore that the unbound macrostate, while energetically disfavored, is entropically stabilized relative to the bound macrostates, leading to approximately equal populations at this λ value.

As shown in Figure 8 for ligand 2, the relatively sharp population switch from the unbound macrostate to the bound macrostate as a function of λ is indeed characteristic of thermodynamic transitions involving large energy/entropy compensation. The equilibrium between unbound and bound macrostates can be described as a pseudo-order/disorder phase transition similar to those observed in protein folding, in which the ordered phase (the bound state) is increasingly destabilized relative to the disordered phase (the unbound state) as λ is decreased and receptor–ligand interactions are turned off. Using a formalism from the protein folding realm, the free energy difference between the unbound and bound macrostate at the half point of the transition $\lambda = \lambda_{1/2}$ (corresponding to the “melting temperature” of protein folding equilibria) is zero, and the steepness of the “melting curve”, shown in Figure 8 for ligands 2 and 9, is proportional to the average binding energy difference between the bound and unbound states. Specifically, as shown in the Appendix, at the half point $\lambda = \lambda_{1/2}$, we have

$$\left(\frac{dP_\lambda(B)}{d\lambda}\right)_{\lambda_{1/2}} = -\frac{\Delta u_{1/2}}{4k_B T} \quad (11)$$

where $P_\lambda(B)$ is the λ -dependent population of the bound state and $\Delta u_{1/2} = \langle u \rangle_{\lambda_{1/2}, B} - \langle u \rangle_{\lambda_{1/2}, U}$ is the difference of the average binding energies of the bound and unbound states at the half point. A very similar relationship exists for folding/unfolding equilibrium as a function of the temperature.⁷⁰ Consequently, the steepness of the unbinding curve at the half point is related to

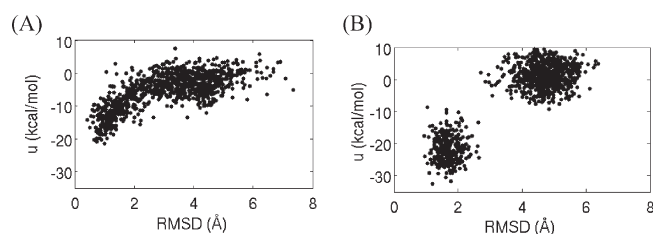


Figure 7. The binding energies vs RMSD of the core region of the ligands relative to the corresponding crystal structures for the conformations of the complex with ligand 2 at $\lambda = 0.5$ (A) and with ligand 20 at $\lambda = 0.65$ (B).

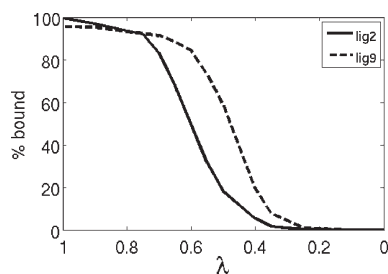


Figure 8. Fractional population of the bound macrostate as a function of λ for the complexes with ligand 2 and ligand 9.

the magnitude of the entropic and energetic changes during the transition, which are exactly equal in magnitude and of opposite sign given that at the half point the binding free energy is zero. The larger these changes, the sharper is the transition from the bound state to the unbound state. Therefore, for ligands that incur large energetic (and entropic) changes in going from the bound to the unbound conformations there is a small range of λ values at which the bound and unbound states are in equilibrium. As further discussed below, this is a crucial aspect to understanding the rate of convergence of binding free energy calculations of this kind as critical interconversions between bound and unbound states occur only in the narrow range of λ in which bound and unbound states are in equilibrium.

From the spread along the y axis of Figure 7, we can see that the distribution of binding energies for ligand 2 at $\lambda = 0.5$ is composed of two distinct and equally important contributions, one from the bound macrostate at low binding energies and the other from the unbound macrostate at less favorable binding energies. To achieve the correct proportions of these two contributions and ultimately reach convergence of the corresponding binding energy distributions, it is necessary to converge the relative populations of the bound and unbound macrostates. Our results show that difficulties of achieving equilibration of this binding/unbinding process is the cause for the overall slow convergence of the BEDAM binding free energy. To see why this is the case, it is useful to analyze the conformational trajectories of HREM replicas. Figure 9 shows the time evolution of the binding energy for the 15 replicas of ligand 2 in a 1 ns section of the BEDAM simulation. Note that in these trajectories λ is not constant, as this is the quantity that is being periodically exchanged with other replicas. We see that during this time replica 1 is trapped at low binding energies at or near the bound state, while other replicas (replicas 2 and 4, for example) remain in the unbound state at high binding energies. Infrequently, some

replicas transition from the bound state to the unbound state or vice versa. For example, replica 3 does so at 0.5 ns, and replica 5 exhibits two short excursions to the unbound state at 0.1 and 0.75 ns. These transitions occur only in a relatively narrow range of λ centered at 0.5, which is the range in which both states have significant populations (see Figure 7). It is the rate of binding/unbinding transitions that determines the convergence of the relative populations of the bound and unbound states and, as discussed above, the convergence of the intermediate binding energy distributions and, ultimately, of the BEDAM binding free energy.

On the basis of these insights, it becomes clear why we were unable to reach convergence for ligand 20 using the same definition of the binding site volume employed for the other ligands. Unlike ligand 2 and the other smaller ligands, none of the replicas of this ligand exhibit binding/unbinding transitions during the BEDAM simulation. Some of the replicas of ligand 20 are trapped in the bound state like replica 1 for ligand 2, and others are confined to the unbound state similarly to replica 2 of ligand 2 (see Figure 9). None of the replicas of ligand 20 exhibited binding/unbinding transitions as, for example, replica 3 of ligand 2. As discussed above, it is not possible to arrive at a converged estimate of the binding free energy without proper equilibration between the bound and unbound states. One likely cause for the lack of transitions is the large binding energy difference between bound and unbound states of ligand 20. See for example the spread in binding energies in Figure 7B. The data shown in Figure 7B for ligand 20 correspond to $\lambda_{1/2} = 0.65$, the value at which we measure equal populations of bound and unbound conformations. However, note that this is only a rough estimate of $\lambda_{1/2}$ for ligand 20 because, lacking binding/unbinding transitions, the bound and unbound states have not reached equilibrium. Nevertheless, these data strongly suggest that the transition binding energy $\Delta u_{1/2}$ for ligand 20 is much larger than that for the other ligands. On the basis of eq 11, we conclude that ligand 20 undergoes binding/unbinding transitions in a much narrower range of λ 's than the other ligands and that, therefore, there is a smaller likelihood that a sufficient number of replicas visit this narrow range for a sufficient length of time to observe transitions.

In addition to the larger binding energy difference between the bound and unbound states, the data in Figure 7 also clearly show that, unlike ligand 2, ligand 20 never visits the conformational space in between the bound and unbound states. (Compare the lack of dots in between the clouds corresponding to the bound and unbound states of ligand 20 in Figure 7B to the nearly continuous density of dots connecting the same states of ligand 2 in Figure 7A.) It must be concluded therefore that a large free energy barrier separates bound and unbound conformations of ligand 20 and that an additional cause of the lack of observed transitions is the slow rate of binding/unbinding transitions even at those value of λ 's where the rate of transitions is maximal. What is the molecular nature of these slow transitions? A recent computational study by Olivieri and Gardebien⁷¹ has confirmed the hypothesis that large conformational change of the so-called 80s loop facilitates the entry and exit of ligands from the FKBP12 binding site.^{72,73} Without this rearrangement, the entryway to the binding site is too constricted to allow a rigid docking binding mechanism. Given that the backbone conformation FKBP12 is harmonically restrained (see Methods), in our calculations this gating movement of the 80s loop cannot occur, thereby hampering binding/unbinding transitions. We were able to observe a

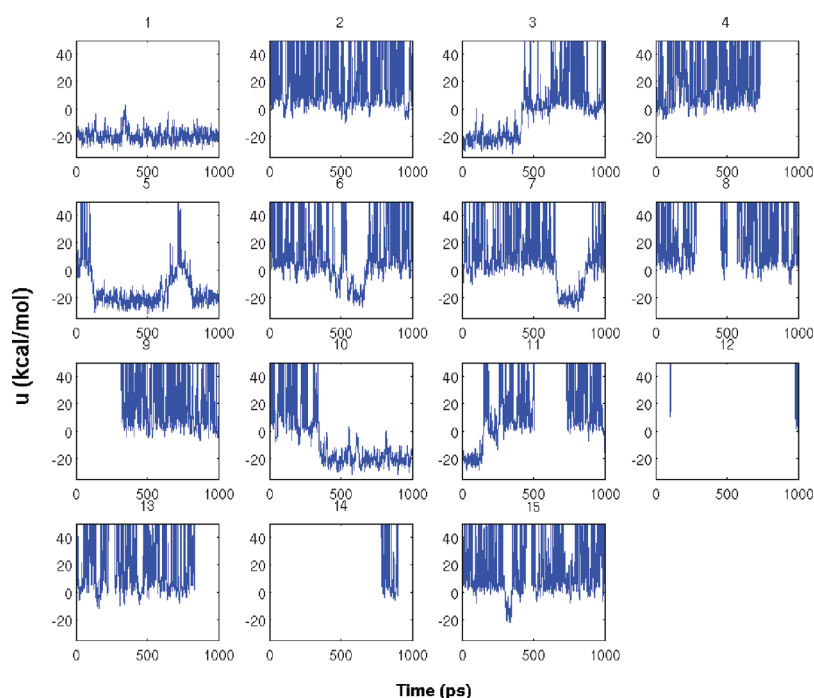


Figure 9. Time evolution of the binding energies of the 15 replicas of the HREMD simulation of the complex with ligand 2. Each panel corresponds to a different replica.

significant number of transitions for the smaller ligands 2–9 because the λ -based alchemical path employed in BEDAM accelerates binding/unbinding transitions beyond what is achievable under physical conditions. However, evidently, this strategy is insufficient for ligand 20, which, given its size and rigidity, has greater difficulty to bind and unbind without receptor rearrangements.

The role of ligand conformational flexibility in the binding mechanism is illustrated in Figure 6, which shows a typical pathway for binding for ligands 2–9. This pathway goes through an intermediate state (panel B) in which only one of the two key receptor–ligand hydrogen bonds is formed (the one between O1 and the side chain of Tyr82), whereas the second hydrogen bond between the ester carbonyl of the ligand and the backbone of Ala56 is not formed because the relevant ligand side chain is rotated away from the donating receptor group (Figure 6, panel B). The last step in the binding transitions consists of the rotation of the ester ligand side chain and the formation of the second hydrogen bond (Figure 6, panel C). Figure 10 indeed shows that rotation of the ligand ester side chain is highly correlated to achieving the crystallographic bound conformation. The majority of the recorded binding events involve this ligand side chain rotation, indicating its likely role in helping the ligand cross the constriction for entering the binding site and, subsequently, form the second receptor–ligand hydrogen bond. Conversely, and unlike the smaller ligands, ligand 20, likely due to its cyclic structure, does not undergo rotation of the ester side chain (Figure 10), and it is forced to follow a less favorable pathway involving the simultaneous formation of both receptor–ligand hydrogen bonds. We believe that the combination of all of these factors—a lack of receptor rearrangement and ligand size and rigidity—is the cause of the lack of observed binding/unbinding transitions for ligand 20 and

the failure to converge its binding free energy using the same protocol used for the other ligands.

To further confirm this hypothesis, we have conducted a BEDAM calculation of ligand 20 using a stricter definition of the complexed state (see Methods) which limits not only the position of the ligand relative to the receptor but also its orientation. This is in principle a valid approach as long as the stricter definition includes all significantly populated conformations of the complex.^{10,58} We have confirmed that this is the case based on the $\lambda = 1$ ensembles of the complexes obtained with the larger binding site definition, which however, as discussed above, resulted in lack of convergence for ligand 20. The purpose of this calculation was to confirm whether circumventing the need to go through the free energy barrier between bound and unbound conformations of the complex with ligand 20 (see Figure 7B) would lead to better convergence of the BEDAM binding free energy estimate. Indeed, with the stricter binding site definition, we observed several transitions between high and low binding energy for ligand 20 and vice versa, similar to those observed for ligand 2 with the larger binding site definition (Figure 9). λ trajectories (Figure 5) also showed more thorough mixing. The larger number of observed binding/unbinding transitions is a consequence of the fact that with the stricter binding site definition unbound conformations of the complex (those with unfavorable binding energies) are conformationally similar to the bound conformations. Because ligand orientations are restrained around the crystallographic pose, the cluster of conformations at high RMSD in Figure 7B is no longer present, and both bound and unbound conformations are found at low RMSDs in such a way that their interconversion no longer requires crossing the free energy barrier. These observations indicate that the binding free energy estimate for ligand 20 reached reasonable convergence with the stricter binding site definition. On the basis of the conformational analysis

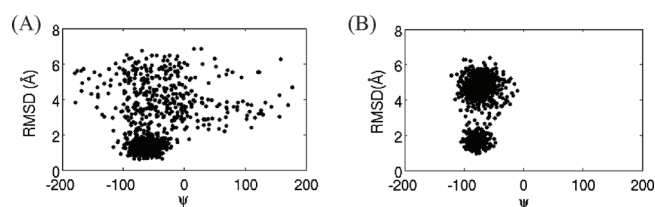


Figure 10. Representation of the reduction in intramolecular conformational freedom of the ligands (ligand 2 in panel A and ligand 20 in panel B) as they bind the receptor. The x axis reports the dihedral angle formed by the C1–C2–C3–O2 atoms of the ligand corresponding to the orientation of one of the two carbonyl groups of the ligand (see text). The y axis reports the RMSD of the core of the ligand relative to the bound crystal structure which serves here as a binding progress coordinate.

of the $\lambda = 1$ trajectories with the original and orientationally restricted binding site definitions (see above), we also expect that this new estimate also retains good accuracy. Although full confirmation of this hypothesis would require further calculations with a range of binding sites definitions, it is reassuring that the resulting binding free energy estimate for ligand 20 is in relatively good agreement with the experimental affinity (Table 1).

CONCLUSIONS

In this work, we have presented the results of BEDAM binding free energy calculations for a series of complexes of the FKBP12 receptor. Analysis of the data generated by parallel HREM simulations have provided valuable insights into the factors that affect convergence of BEDAM binding free energy calculations and, importantly, how to detect and analyze these factors. We have shown that the BEDAM protocol is applicable to protein–ligand systems of the size and complexity often found in pharmaceutical applications. We found that reasonable convergence of the calculations can be achieved for these systems if the λ schedule is adjusted appropriately to ensure sufficient overlaps between neighboring binding energy distributions and if a sufficient number of conformational transitions occur so as to achieve equilibration between the bound and unbound macrostates of the complex. Because of the latter, often underappreciated, requirement, increasing the number of replicas improves the binding energy distributions overlap, but it does not necessarily lead to an improvement of the convergence rate of the binding free energy.

The BEDAM protocol employed here does not rely on the exact structural knowledge of the binding mode; it employs a minimally restrained conformational sampling protocol which is capable of identifying potential multiple bound poses of the complex. The good correspondence between the bound ensemble and the crystallographic bound structure in our simulation is solely due to the ability of the HREM sampling protocol to explore many conformations and the ability of the energy function to recognize conformations similar to the crystallographic structure as more favorable. By performing binding free energy calculations in an unrestrained fashion, we were able to resolve a common binding pathway which involved the step-wise formation of two hydrogen bonds between the ligand and receptor (see Figure 6). We are also exploring the possibility of using BEDAM HREM data as input for the construction of network models of binding similar to the use of temperature

replica exchange data to construct network models for protein folding and dynamics.^{74,75}

However, as we have shown, our model, which allows replicas to visit both bound and unbound conformations of the complex, also requires that independent transitions be observed between these two states in order to reach convergence of the binding free energy. We showed that these transitions have the features of pseudo-order/disorder phase transition analogous to protein folding equilibria. The lack of a sufficient number of transitions prevented convergence for ligand 20 unless a more restrained setup was used. We are currently evaluating ways to increase the occurrence of binding/unbinding transitions with the aim of improving convergence even in cases when the model of the complex explores a large variety of conformations and exhibits phase change characteristics.⁷⁶

Even though they are difficult to model, we have shown that to achieve a realistic representation of the binding equilibrium it is necessary to include the role of conformational entropy loss and intramolecular energetic strain. Although in this system increased affinity qualitatively tracks more favorable receptor–ligand interactions (Table 1), neglecting reorganization free energies grossly overestimates variations of binding affinities from one ligand to another.

Having included in the model the main thermodynamic driving forces and having achieved a good quality of conformational sampling and convergence in the numerical calculations, we believe that the level of agreement of the binding free energy estimates with the experimental affinities primarily reflects the accuracy of the potential energy model (here OPLS-AA with AGBNP2 implicit solvation). The model reproduces ligand ranking reasonably well, and we have shown that quantitative agreement can be achieved by adopting a more realistic geometrical model of the solvent accessible surface area of the complex. These promising results indicate that the foundations of the OPLS-AA/AGBNP2 effective potential are solid but that further development and parametrization is required to achieve improved accuracy and transferability in binding free energy applications.

APPENDIX

In this Appendix, we derive eq 11. The population $P_\lambda(B)$ of the bound state B at λ is given by

$$P_\lambda(B) = \frac{1}{Z_\lambda} \int e^{-\beta[V_0(\mathbf{r}) + \lambda u(\mathbf{r})]} \Theta_B(\mathbf{r}) \, d\mathbf{r} = \frac{Z_\lambda(B)}{Z_\lambda} \quad (12)$$

where integration is over all of the possible conformations of the complex, the λ -dependent potential energy is from eq 4, $\Theta_B(\mathbf{r})$ is an indicator function equal to 1 if conformation r belongs to the bound macrostate and 0 otherwise, Z_λ is the configurational partition function of the complex at λ , and $Z_\lambda(B)$ is the configurational partition function of only the bound macrostate. Differentiation of eq 12 leads to the expression

$$\frac{dP_\lambda(B)}{d\lambda} = -\beta P_\lambda(B) (\langle u \rangle_{\lambda,B} - \langle u \rangle_\lambda) \quad (13)$$

where the first term in parentheses comes from the differentiation of $Z_\lambda(B)$ and the second from the differentiation of Z_λ at the

denominator of eq 12

$$\langle u \rangle_{\lambda} = \frac{1}{Z_{\lambda}(B)} \int u(\mathbf{r}) e^{-\beta[V_0(\mathbf{r}) + \lambda u(\mathbf{r})]} d\mathbf{r} \quad (14)$$

is the average binding energy at λ and

$$\langle u \rangle_{\lambda, B} = \frac{1}{Z_{\lambda}(B)} \int u(\mathbf{r}) e^{-\beta[V_0(\mathbf{r}) + \lambda u(\mathbf{r})]} \Theta_B(\mathbf{r}) d\mathbf{r} \quad (15)$$

is the average binding energy of the bound macrostate at λ . The average binding energy $\langle u \rangle_{\lambda}$ is the weighted sum of the average binding energies $\langle u \rangle_{\lambda, B}$ and $\langle u \rangle_{\lambda, U}$, respectively, in the bound (B) and unbound (U) macrostates (which together are assumed to comprise all of the conformations of the complex)

$$\langle u \rangle_{\lambda} = P_{\lambda}(B) \langle u \rangle_{\lambda, B} + P_{\lambda}(U) \langle u \rangle_{\lambda, U} \quad (16)$$

At $\lambda = \lambda_{1/2}$ where $P_{\lambda}(B) = P_{\lambda}(U) = 1/2$, we have

$$\langle u \rangle_{\lambda_{1/2}} = \frac{1}{2} (\langle u \rangle_{\lambda_{1/2}, B} + \langle u \rangle_{\lambda_{1/2}, U}) \quad (17)$$

which, when substituted into eq 13, yields

$$\left(\frac{dP_{\lambda}(B)}{d\lambda} \right)_{\lambda=\lambda_{1/2}} = -\frac{\beta}{4} (\langle u \rangle_{\lambda_{1/2}, B} - \langle u \rangle_{\lambda_{1/2}, U}) = -\frac{1}{k_B T} \Delta u_{1/2} \quad (18)$$

which is eq 11.

AUTHOR INFORMATION

Corresponding Authors

*E-mail: emilio@biomaps.rutgers.edu; ronlevy@lutece.rutgers.edu.

Note

The authors declare no competing financial interest.

ACKNOWLEDGMENT

This work has been supported in part by a research grant from the National Institutes of Health (GM30580). The calculations reported in this work have been performed at the BioMaPS High Performance Computing Center at Rutgers University, funded in part by the NIH shared instrumentation grants no. 1 S10 RR022375 and 1 S10 RR027444, and on the Lonestar4 cluster at the Texas Advanced Computing Center under TeraGrid/XSEDE National Science Foundation allocation grant no. TG-MCB100145.

REFERENCES

- Jorgensen, W. L. *Nature* **2010**, *466*, 42–43.
- Gallicchio, E.; Levy, R. M. *Curr. Opin. Struct. Biol.* **2011**, *21*, 161–166.
- Chodera, J. D.; Mobley, D. L.; Shirts, M. R.; Dixon, R. W.; Branson, K.; Pande, V. S. *Curr. Opin. Struct. Biol.* **2011**, *21*, 150–160.
- Shoichet, B. K.; McGovern, S. L.; Wei, B.; Irwin, J. J. *Curr. Opin. Chem. Biol.* **2002**, *6*, 439–446.
- Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R. *J. Med. Chem.* **2006**, *49*, 534–553.
- Zhou, Z.; Felts, A. K.; Friesner, R. A.; Levy, R. M. *J. Chem. Inf. Model.* **2007**, *47*, 1599–1608.
- Trott, O.; Olson, A. J. *J. Comput. Chem.* **2010**, *31*, 455–461.
- Guvench, O.; MacKerell, A. D. *Curr. Opin. Struct. Biol.* **2009**, *19*, 56–61.
- Gilson, M. K.; Zhou, H.-X. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21–42.
- Gallicchio, E.; Levy, R. M. In *Advances in Protein Chemistry and Structural Biology*; Christov, C., Ed.; Academic Press: New York, 2011; Vol. 85, Chapter Recent Theoretical and Computational Advances for Modeling Protein-Ligand Binding Affinities, pp 27–80.
- Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. *Biophys. J.* **1997**, *72*, 1047–1069.
- Hansson, T.; Marelus, J.; Aqvist, J. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 27–35.
- Zhou, R.; Friesner, R. A.; Ghosh, A.; Rizzo, R. C.; Jorgensen, W. L.; Levy, R. M. *J. Phys. Chem.* **2001**, *105*, 10388–10397.
- Su, Y.; Gallicchio, E.; Das, K.; Arnold, E.; Levy, R. *J. Chem. Theory Comput.* **2007**, *3*, 256–277.
- Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. *Acc. Chem. Res.* **2000**, *33*, 889–897.
- Chang, C.-E.; Gilson, M. K. *J. Am. Chem. Soc.* **2004**, *126*, 13156–13164.
- Mobley, D. L.; Dill, K. A. *Structure* **2009**, *17*, 489–498.
- Oostenbrink, C.; van Gunsteren, W. F. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 6750–6754.
- Chipot, C.; Pohorille, A. *Free Energy Calculations. Theory and Applications in Chemistry and Biology*; Springer Series in Chemical Physics; Springer: Berlin, 2007.
- Knight, J. L.; Brooks, C. L. *J. Comput. Chem.* **2009**, *30*, 1692–1700.
- Michel, J.; Essex, J. W. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 639–658.
- Mobley, D. L.; Chodera, J. D.; Dill, K. A. *J. Chem. Theory Comput.* **2007**, *3*, 1231–1235.
- Ge, X.; Roux, B. *J. Phys. Chem. B* **2010**, *114*, 9525–9539.
- Colizzi, F.; Perozzo, R.; Scapoza, L.; Recanatini, M.; Cavalli, A. *J. Am. Chem. Soc.* **2010**, *132*, 7361–71.
- Miyata, T.; Ikuta, Y.; Hirata, F. *J. Chem. Phys.* **2010**, *133*, 044114.
- Tembe, B. L.; McCammon, J. A. *Comput. Chem.* **1984**, *8*, 281.
- Jorgensen, W. L.; Buckner, J. K.; Boudon, S.; Tirado-Rives, J. *J. Chem. Phys.* **1988**, *89*, 3742.
- Mobley, D. L.; Graves, A. P.; Chodera, J. D.; McReynolds, A. C.; Shoichet, B. K.; Dill, K. A. *J. Mol. Biol.* **2007**, *371*, 1118–1134.
- Gallicchio, E.; Lapelosa, M.; Levy, R. M. *J. Chem. Theory Comput.* **2010**, *6*, 2961–2977.
- Woods, C. J.; Essex, J. W.; King, M. A. *J. Phys. Chem. B* **2003**, *107*, 13703–13710.
- Jiang, W.; Roux, B. *J. Chem. Theory Comput.* **2010**, *6*, 2559–2565.
- Holt, D. A.; Luengo, J. I.; Yamashita, D. S.; Oh, H. J.; Konialian, A. L.; Yen, H. K.; Rozamus, L. W.; Brandt, M.; Bossard, M. J. *J. Am. Chem. Soc.* **1993**, *115*, 9925–9938.
- Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011–1021.
- Gallicchio, E.; Andrec, M.; Felts, A. K.; Levy, R. M. *J. Phys. Chem. B* **2005**, *109*, 6722–6731.
- Shirts, M. R.; Chodera, J. D. *J. Chem. Phys.* **2008**, *129*, 124105.
- Gallicchio, E.; Levy, R. *J. Comput. Chem.* **2004**, *25*, 479–499.
- Gallicchio, E.; Paris, K.; Levy, R. M. *J. Chem. Theory Comput.* **2009**, *5*, 2544–2564.
- Bossard, M. J.; Bergsma, D. J.; Brandt, M.; Livi, G. P.; Eng, W. K.; Johnson, R. K.; Levy, R. M. *Biochem. J.* **1994**, *297* (Pt 2), 365–372.
- Wang, J.; Deng, Y.; Roux, B. *Biophys. J.* **2006**, *91*, 2798–2814.
- Fujitani, H.; Tanida, Y.; Ito, M.; Jayachandran, G.; Snow, C. D.; Shirts, M. R.; Sorin, E. J.; Pande, V. S. *J. Chem. Phys.* **2005**, *123*, 084108.
- Hajduk, P. J.; Greer, J. *Nat. Rev. Drug Discovery* **2007**, *6*, 211–219.
- Congreve, M.; Chessari, G.; Tisi, D.; Woodhead, A. J. *J. Med. Chem.* **2008**, *51*, 3661–3680.
- Newman, J.; Fazio, V. J.; Caradoc-Davies, T. T.; Branson, K.; Peat, T. S. *J. Biomol. Screen.* **2009**, *14*, 1245–1250.
- Deng, Y.; Roux, B. *J. Phys. Chem. B* **2009**, *113*, 2234–2246.

- (45) Csermely, P.; Palotai, R.; Nussinov, R. *Trends Biochem. Sci.* **2010**, *35*, 539–546.
- (46) Gao, C.; Park, M.-S.; Stern, H. A. *Biophys. J.* **2010**, *98*, 901–910.
- (47) Lapelosa, M.; Gallicchio, E.; Arnold, G. F.; Arnold, E.; Levy, R. M. *J. Mol. Biol.* **2009**, *385*, 675–691.
- (48) Lapelosa, M.; Arnold, G. F.; Gallicchio, E.; Arnold, E.; Levy, R. M. *J. Mol. Biol.* **2010**, *397*, 752–766.
- (49) Bizzarri, M.; Marsili, S.; Procacci, P. *J. Phys. Chem. B* **2011**, *115*, 6193–6201.
- (50) Bizzarri, M.; Tenori, E.; Martina, M. R.; Marsili, S.; Caminati, G.; Menichetti, S.; Procacci, P. *J. Phys. Chem. Lett.* **2011**, *2*, 2834–2839.
- (51) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.
- (52) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.
- (53) Wei, B. Q.; Baase, W. A.; Weaver, L. H.; Matthews, B. W.; Shoichet, B. K. *J. Mol. Biol.* **2002**, *322*, 339–355.
- (54) Buelens, F. P.; Grubmüller, H. *J. Comput. Chem.* **2012**, *33*, 25–33.
- (55) Tan, Z. *J. Am. Stat. Assoc.* **2004**, *99*, 1027–1036.
- (56) Chernick, M. R. *Bootstrap Methods: A Guide for Practitioners and Researchers*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, 2008.
- (57) Banks, J. L.; et al. *J. Comput. Chem.* **2005**, *26*, 1752–1780.
- (58) Mihailescu, M.; Gilson, M. K. *Biophys. J.* **2004**, *87*, 23–36.
- (59) Borech, S.; Tettinger, F.; Leitgeb, M.; Karplus, M. *J. Phys. Chem. B* **2003**, *107*, 9535–9551.
- (60) Nicholls, A.; Sharp, K. A.; Honig, B. *Proteins* **1991**, *11*, 281–96.
- (61) Gallicchio, E.; Kubo, M. M.; Levy, R. M. *J. Phys. Chem. B* **2000**, *104*, 6271–6285.
- (62) Zhou, H.-X.; Gilson, M. K. *Chem. Rev.* **2009**, *109*, 4092–4107.
- (63) Pohorille, A.; Jarzynski, C.; Chipot, C. *J. Phys. Chem. B* **2010**, *114*, 10235–10253.
- (64) Zheng, W.; Andrec, M.; Gallicchio, E.; Levy, R. M. *J. Phys. Chem. B* **2008**, *112*, 6083–6093.
- (65) Meng, Y.; Roitberg, A. E. *J. Chem. Theory Comput.* **2010**, *6*, 1401–1412.
- (66) Mitsutake, A.; Mori, Y.; Okamoto, Y. *Physics Procedia* **2010**, *4*, 89–105.
- (67) Wang, L.; Friesner, R. A.; Berne, B. J. *J. Phys. Chem. B* **2011**, *115*, 9431–9438.
- (68) Khavrutskii, I. V.; Wallqvist, A. *J. Chem. Theory Comput.* **2011**, *7*, 3001–3011.
- (69) Min, D.; Chen, M.; Zheng, L.; Jin, Y.; Schwartz, M. A.; Sang, Q.-X. A.; Yang, W. *J. Phys. Chem. B* **2011**, *115*, 3924–3935.
- (70) Finkelstein, A. V.; Ptitsyn, O. *Protein Physics*; Academic Press: San Diego, CA, 2002.
- (71) Olivieri, L.; Gardebien, F. *J. Chem. Theory Comput.* **2011**, *7*, 725–741.
- (72) Ivery, M. T.; Weiler, L. *Bioorg. Med. Chem.* **1997**, *5*, 217–32.
- (73) Wilson, K. P.; Yamashita, M. M.; Sintchak, M. D.; Rotstein, S. H.; Murcko, M. A.; Boger, J.; Thomson, J. A.; Fitzgibbon, M. J.; Black, J. R.; Navia, M. A. *Acta Crystallogr., Sect. D* **1995**, *51*, 511–521.
- (74) Zheng, W.; Gallicchio, E.; Deng, N.; Andrec, M.; Levy, R. M. *J. Phys. Chem. B* **2011**, *115*, 1512–1523.
- (75) Deng, N.; Zheng, W.; Gallicchio, E.; Levy, R. *J. Am. Chem. Soc.* **2011**, *133*, 9387–9894.
- (76) Kim, J.; Straub, J. E. *J. Chem. Phys.* **2010**, *133*, 154101.

Force Field Benchmark of Organic Liquids: Density, Enthalpy of Vaporization, Heat Capacities, Surface Tension, Isothermal Compressibility, Volumetric Expansion Coefficient, and Dielectric Constant

Carl Caleman,[†] Paul J. van Maaren,[‡] Minyan Hong,[‡] Jochen S. Hub,[‡] Luciano T. Costa,[§] and David van der Spoel^{*,†}

[†]Center for Free-Electron Laser Science, Deutsches Elektronen-Synchrotron Notkestraße 85, DE-22607 Hamburg, Germany

[‡]Department of Cell and Molecular Biology, Uppsala University, Husargatan 3, Box 596, SE-75124 Uppsala, Sweden

[§]Departamento de Ciências Exatas, Federal University of Alfenas—MG Rua Gabriel Monteiro da Silva, 700 Alfenas—MG CEP: 37130-000, Brazil

S Supporting Information

ABSTRACT: The chemical composition of small organic molecules is often very similar to amino acid side chains or the bases in nucleic acids, and hence there is no a priori reason why a molecular mechanics force field could not describe both organic liquids and biomolecules with a single parameter set. Here, we devise a benchmark for force fields in order to test the ability of existing force fields to reproduce some key properties of organic liquids, namely, the density, enthalpy of vaporization, the surface tension, the heat capacity at constant volume and pressure, the isothermal compressibility, the volumetric expansion coefficient, and the static dielectric constant. Well over 1200 experimental measurements were used for comparison to the simulations of 146 organic liquids. Novel polynomial interpolations of the dielectric constant (32 molecules), heat capacity at constant pressure (three molecules), and the isothermal compressibility (53 molecules) as a function of the temperature have been made, based on experimental data, in order to be able to compare simulation results to them. To compute the heat capacities, we applied the two phase thermodynamics method (Lin et al. *J. Chem. Phys.* **2003**, *119*, 11792), which allows one to compute thermodynamic properties on the basis of the density of states as derived from the velocity autocorrelation function. The method is implemented in a new utility within the GROMACS molecular simulation package, named `g_dos`, and a detailed exposé of the underlying equations is presented. The purpose of this work is to establish the state of the art of two popular force fields, OPLS/AA (all-atom optimized potential for liquid simulation) and GAFF (generalized Amber force field), to find common bottlenecks, i.e., particularly difficult molecules, and to serve as a reference point for future force field development. To make for a fair playing field, all molecules were evaluated with the same parameter settings, such as thermostats and barostats, treatment of electrostatic interactions, and system size (1000 molecules). The densities and enthalpy of vaporization from an independent data set based on simulations using the CHARMM General Force Field (CGenFF) presented by Vanommeslaeghe et al. (*J. Comput. Chem.* **2010**, *31*, 671) are included for comparison. We find that, overall, the OPLS/AA force field performs somewhat better than GAFF, but there are significant issues with reproduction of the surface tension and dielectric constants for both force fields.

1. INTRODUCTION

Parameters in most force fields have been derived incrementally, that is, building on previous work by adding support for different chemical moieties in a sequential fashion. While the focus of many force fields is on biomolecules, the chemical basis lies in organic molecules. Of the major force fields available today OPLS/AA (optimized parameters for liquid simulations, all atoms) is one of the few that “specializes” in simple liquids.¹ The generalized Amber force field (GAFF) was introduced recently² (together with the Antechamber set of programs³) to aid in the derivation of force field parameters for small molecules that are often involved in binding to biomolecules. Accurate parameters are crucial for predicting, for instance, the Gibbs energy of ligand binding, a key property in drug design.⁴ The GAFF parameters for small molecules are intended to be combined with the Amber force field⁵ although there are studies of proteins using GAFF parameters as well.⁶

A critical component in force field development is generation of partial charges. The method for deriving partial charges by

optimizing their values to reproduce the electrostatic potential (ESP) was introduced in the 1980s by Kollman et al.^{7,8} The electron density taken from a quantum chemistry calculation, together with the nuclear charges, generates an electrostatic potential around the molecule. Typically, the set of partial charges for a molecule, for use in force field calculations, is determined by minimizing the (square) difference between the ESP generated by the partial charges and the ESP generated by the quantum chemistry calculation. A set of partial charges (or indeed any atom-centered set of spherically distributed charges) can never completely reproduce the ESP due to the fact that electron density is not completely spherically symmetric around the nuclei (for instance, due to p and higher orbitals). A further issue is due to the fact that the fitting points are highly correlated, and hence atoms far from the ESP data points (e.g., the buried

Received: October 18, 2011

Published: December 07, 2011

carbon in isobutanol) may end up being a sink for the fit^{9,10} and get arbitrary values. An ad hoc refinement of the ESP method to overcome this problem is the restrained ESP (RESP) method.¹¹ The RESP method does the same fit, however with an added penalty on the absolute value of the charge. The RESP method is an integral part of the Antechamber package,^{2,3} which relies on either quantum calculations or empirical methods, such as AM1-BCC,^{12,13} to provide the partial charges.

Mobley et al. tested the performance of GAFF parameters for Gibbs energies of hydration using two different water models.^{14,15} They paid particular attention to the way the partial charges were determined and found that the final results are related to the level of theory used, something that was corroborated by Wallin et al., who did a similar study of charge schemes for ligand binding.¹⁶ The CM1 charge model for OPLS/AA,¹⁷ used in the study of Wallin et al.,¹⁶ performs well,^{18,19} although some degradation for conformational energetics is expected. The differences are generally considered to be minor.¹ There are some drawbacks with these studies however. First, they involve complex systems, where a subset of the parameters was changed and the “quality” of the charges evaluated on the basis of a single number, the free energy, hereby ignoring the interdependency between Lennard-Jones parameters and point charges. Second, free energy calculations depend critically on the amount of the sampling that was used, although it is possible to ascertain that the errors due to sampling are small.²⁰ In order to test the validity of force field, it would be good to take one step back and evaluate the performance for simple systems first, in order to avoid systematic errors due to water model and/or protein force fields. A recent review by Jorgensen and Tirado-Rives provides further background information on the topic of force field development.¹

To assess the state of the art of GAFF and OPLS/AA force fields, we provide a comprehensive benchmark of the liquid properties of molecules in each of the GAFF and OPLS/AA force fields. Previous simulations of mixtures of alcohol and water^{21,22} using the OPLS/AA force field showed that many properties of the pure liquids are reproduced faithfully, but the heat of mixing and the density of mixing are slightly, but significantly, off. Similar comparisons of force fields for water models are numerous in the literature (see for example, refs 23–29), while for organic liquids there are some papers by Kaminski and Jorgensen,^{30,31} and a recent paper by Wang and Tingjun,³² which we discuss in the Discussion section.

Liquid properties are usually known experimentally with high accuracy, and their calculation is most often straightforward. Rather, the time goes into the preparation and equilibration of the systems. A total of 146 molecular liquids was prepared and simulated using these force fields in the GROMACS molecular simulation package,^{33–36} and from these molecular dynamics simulations, we extract the density ρ (from constant pressure simulations), the enthalpy of vaporization ΔH_{vap} , the heat capacities at constant pressure c_p and volume c_v , the volumetric expansion coefficient α_p , the isothermal compressibility κ_T , the surface tension γ , and the static dielectric constant $\epsilon(0)$. Although, in principle, more observables could be computed, this set includes the most important thermodynamic properties of the liquids, including temperature derivatives of energy and volume. The intention of this work is to supply a large number of tests for further force field development. To this end, the topologies and structures have been made available on a dedicated Web site at <http://virtualchemistry.org>, while the simulation parameters are available as Supporting Information to this paper. These topologies and structure files may be useful for simulations of

biomolecules in organic liquids as well. The recently presented all atom CHARMM general force field (CGenFF)³⁷ would be an equally well suited candidate for inclusion in this comparison, but we have chosen to limit our simulations to two force fields only. However, to allow the reader to compare OPLS/AA and GAFF to a similar study based on CGenFF, we have included results on density and enthalpy of vaporization from that paper.³⁷

2. METHODS

2.1. Energy Function. Most force fields use the same functional form for the intermolecular part of the interaction function, based on the Coulomb potential and the Lennard-Jones potential:

$$V_{\text{nb}}(r_{ij}) = \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} + 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] \quad (1)$$

where r_{ij} is the distance between two atoms i and j , q_i and q_j are the partial charges on the atoms, ϵ_0 is the permittivity of vacuum, σ_{ij} is the van der Waals radius, and ϵ_{ij} is the well-depth for this atom pair. In most force fields, the parameters σ_{ij} and ϵ_{ij} are derived from the atomic values σ_i and ϵ_i using a simple equation (the combination rule). Suffice to say that we have applied the standard combination rules for GAFF (Lorentz–Berthelot³⁸) and for OPLS/AA ($\sigma_{ij} = (\sigma_i \sigma_j)^{1/2}$ and $\epsilon_{ij} = (\epsilon_i \epsilon_j)^{1/2}$)³⁹ in this work.

2.2. Molecule Selection and Preparation. A set of organic molecules was selected for which both enthalpy of vaporization and density are known at room temperature. Models for these molecules were built using either PRODRG⁴⁰ or Molden.⁴¹ These molecules were optimized using the Gaussian 03 suite of programs⁴² at the Hartree–Fock level with the 6-311G** basis set.^{43–47}

2.2.1. OPLS/AA Topologies. The OpenBabel (<http://openbabel.org>) code was used to extract a coordinate file including connectivity information from the Gaussian output files, and this file was used to generate an initial topology using the GROMACS tools³⁵ for the OPLS/AA force field.^{1,39} The topologies were checked manually for correctness before using them, making sure that the total charge of the molecule is zero, and also that the atom types were correct. For molecules containing linear groups (e.g., nitriles), a virtual site construction was added to the topologies preserving the moment of inertia and the total mass, in order to keep the groups perfectly linear.⁴⁸

2.2.2. GAFF Topologies. For the simulations where GAFF² was used, the Antechamber software^{2,3} was employed to generate the topologies from the coordinate files (which were generated as explained above). Gaussian 03⁴² at the Hartree–Fock level with the 6-311G** basis set^{43–47} (as provided by the Basis Set Exchange Web site^{49,50}) and Merz–Singh–Kollman (MK) scheme⁷ were used to determine the partial charges in Gaussian. This particular basis set was used because it is very similar to the 6/31G* basis set,⁵¹ which is the default for GAFF, while simultaneously supporting a larger number of elements (e.g., I). The MK radius for I is not implemented in Antechamber, we used $R_I = 2.15$ Å. The `amb2gmx.pl` script⁵² was used to convert the AMBER topologies into the GROMACS format (this script is available online at <http://ffamber.cns.msu.edu/>). The final partial charges were calculated using the RESP method¹¹ as implemented in Antechamber, and we manually checked that the charges were sane. Note that RESP can be used with any QM method producing electrostatics, not just with HF/6-311G**.

Table 1. Simulation Characteristics for the Different Simulation Types

name	length	# molecules	ensemble	constraints	electrostatics
LIQ	10 ns	1000	NPT	all bonds	PME
GAS	100 ns	1	NVT	all bonds	all interactions
SURF	10 ns	1000	NVT	all bonds	PME
DOS	100 ps	1000	NVT	none	PME

No modifications for linear group were made for the GAFF topologies, where the Antechamber software³ generates a near-linear angle term instead.

2.2.3. Liquid Simulation Box Preparation. To generate liquid simulation boxes, we first made a $2 \times 2 \times 2 \text{ nm}^3$ box containing a single molecule. From 125 such single molecule boxes, we built up a $10 \times 10 \times 10 \text{ nm}^3$ box. These boxes were simulated under high pressure (100 bar) to force the molecules into the liquid phase, and finally we let the systems relax under normal pressure (1 bar) to reach an equilibrated system. For the equilibration simulations, we used Berendsen's coupling algorithm⁵³ because of its efficient relaxation properties.⁵⁴ To generate our final simulation boxes, we stacked $2 \times 2 \times 2$ of the 125 molecule boxes and ran an additional equilibration simulation. The absolute drift in total energy was automatically checked in the equilibration and production simulations, and the simulations were continued until the drift was below 0.5 J/mol/ns per degree of freedom, which is a very strict criterion but which is necessary to accurately compute fluctuation properties.

2.3. Simulation Parameters. The GROMACS suite of programs was used for all simulations.^{33–36} Following previous simulations of alcohol water mixtures^{21,22} using the OPLS/AA force field,^{1,39} we employed a 1.1 nm cutoff for Lennard-Jones interactions and the same distance as the switching distance for the particle mesh Ewald (PME) algorithm for computing Coulomb interactions.^{54,55} Although the OPLS/AA force field was not developed for use with PME, extensive studies on water models⁵⁶ and proteins in water⁵⁷ have shown that correspondence of simulation results with experimental data improves considerably when long-range interactions are taken into account explicitly—irrespective of the force field used. Analytic corrections to pressure and potential energies were made to compensate for the truncation of the Lennard-Jones interactions.³⁸ In the production simulations, we used the Nosé–Hoover algorithm for temperature coupling,^{58,59} in order to provide correct fluctuations, which is necessary to compute fluctuation properties. A time constant for coupling of 1 ps (corresponding to a mass parameter Q of 7.6 ps at room temperature) was used, which is in the range of time scales for intermolecular collisions, as recommended by Holian et al.⁶⁰ For production simulations at constant pressure, the Parrinello–Rahman pressure coupling⁶¹ algorithm was used with compressibility set to $5 \times 10^{-5} \text{ bar}^{-1}$ and a time constant of 5 ps. The temperatures of the simulations were selected to fit the experimental data available. In most simulations, the bonds were constrained using the LINCS algorithm^{62,63} for all molecules, applying two iterations in order to obtain good energy conservation. Periodic boundary conditions were used in all liquid phase simulations.

Four types of production run simulations were performed according to Table 1. The density of states (DOS) production simulations were performed under constant volume conditions, but they were preceded by equilibration simulations

under NPT (without constraints) in order to obtain the equilibrium density at $P = 1 \text{ bar}$ for the subsequent DOS simulations. In the DOS simulations, slightly stricter energy conservation parameters were used: a neighbor list buffer of 0.3 nm, combined with a switched Lennard-Jones and short-range electrostatics term (1.0–1.1 nm), see reference 56 for a description of the functional form.

The GAS simulations were done using a stochastic dynamics (SD) integrator, which adds a friction and a noise term to Newton's equation of motion:

$$m_i \frac{d^2 \mathbf{r}_i}{dt^2} = -m_i \xi_i \frac{d\mathbf{r}_i}{dt} + \mathbf{F}_i(\mathbf{r}) + \rho_i \quad (2)$$

where m_i is the mass of atom i , ξ_i is a friction constant, and $\rho(t)$ is a noise process with

$$\langle \rho_i(t) \rho_j(t+s) \rangle = 2m_i \xi_i k_B T \delta(s) \delta_{ij} \quad (3)$$

where k_B is Boltzmann's constant, T is the temperature, $\delta(s)$ is the Dirac δ function, and δ_{ij} is the Kronecker δ function. A leapfrog algorithm adapted for SD simulations⁶⁴ was used to integrate eq 2. When $1/\xi_i$ is large compared to the time scales present in the system, SD functions like molecular dynamics with stochastic temperature-coupling. One of the benefits with SD as compared to MD is that when simulating a system in a vacuum there is no accumulation of errors for the overall translational and rotational degrees of freedom, making sampling of different configuration states more accurate. SURF and LIQ simulations were done using a conventional MD leapfrog integrator.⁶⁵ To enable replication of our simulations and detailed scrutiny of the data, we provide all simulation parameters for each type of run, as well as starting structures and topologies. These files, in GROMACS format, are available for downloading at <http://virtualchemistry.org>.

To ensure that our liquid systems did not freeze during the simulations, we monitored the changes in diffusion constant ΔD as derived from the mean square displacement during the simulations, defined as

$$\Delta D = \frac{2(D_{\text{end}} - D_{\text{begin}})}{D_{\text{end}} + D_{\text{begin}}} \quad (4)$$

The subscript “begin” means the value is an average over the 1000–1500 ps of the simulation, and “end” means over 8500–9000 ps. $|\Delta D|$ is close to zero for most simulations, indicating that D is approximately the same in the beginning and at the end of the simulation. We also verified that $D > 0$ for all simulations. For the simulations where $|\Delta D| \geq 0.5$, we ensured that the systems indeed were not frozen, by inspecting the full mean square displacement curve and the trajectory of the simulations. In the Supporting Information (Figure S1), we show ΔD for all of the liquid simulations.

2.4. Analysis. The density ρ in a constant pressure simulation follows trivially from the mass M of the system divided by the volume V :

$$\rho = \frac{M}{\langle V \rangle} \quad (5)$$

The enthalpy of vaporization can be computed from

$$\Delta H_{\text{vap}} = (E_{\text{intra}}(\text{g}) + k_B T) - (E_{\text{intra}}(\text{l}) + E_{\text{inter}}(\text{l})) \quad (6)$$

where E_{intra} is the intramolecular energy in either the gas (g) phase or the liquid (l) phase and E_{inter} represents the intermolecular energy of the system. In practice, we can simply evaluate

$$\Delta H_{\text{vap}} = (E_{\text{pot}}(\text{g}) + k_{\text{B}}T) - E_{\text{pot}}(\text{l}) \quad (7)$$

ρ was determined from LIQ simulations and ΔH_{vap} from LIQ and GAS simulations.

The SURF simulations were done using liquid boxes, the size of which in the z direction was extended by a factor of 3, generating a simulation box with two liquid–vacuum interfaces. The surface tension γ then follows from

$$\gamma(t) = \frac{L_z}{2} \left(P_z(t) - \frac{P_x(t) + P_y(t)}{2} \right) \quad (8)$$

where P_n is the pressure component in direction n and L_z is the length of the box in the z direction (perpendicular to the surfaces).

Static dielectric constants $\epsilon(0)$ were computed on the basis of the fluctuations of the total dipole moment \mathbf{M} of the simulation box^{66,67} in the LIQ simulations:

$$\epsilon(0) = 1 + \frac{4\pi}{3} \frac{\langle \mathbf{M}^2 \rangle - \langle \mathbf{M} \rangle^2}{Vk_{\text{B}}T} \quad (9)$$

where V is the volume of the simulation box. Errors were estimated by block-averaging over 10 blocks of 1 ns. In order to verify the validity of eq 9, we computed the autocorrelation time τ_{M} of the total dipole moment \mathbf{M} in the simulation boxes (from the integral of the autocorrelation function). In order for fluctuations to be well-defined, τ_{M} should be at least an order of magnitude shorter than the simulation length. Henceforth, we omitted the dielectric constants for those systems where τ_{M} was longer than 1 ns. For those systems where this was the case, longer simulations of 50 ns were performed, in most cases without any improvement.

The fluctuation properties α_{p} (the volumetric thermal expansion coefficient) and κ_{T} (the isothermal compressibility) are computed from the LIQ simulations according to³⁸

$$\langle \delta V \delta H \rangle = k_{\text{B}}T^2 \langle V \rangle \alpha_{\text{p}} \quad (10)$$

where H is the enthalpy and δ indicates the fluctuations, and

$$\langle \delta V^2 \rangle = k_{\text{B}}T \langle V \rangle \kappa_{\text{T}} \quad (11)$$

These two properties can be related to the difference between heat capacities at constant pressure and constant volume through

$$\Delta c = c_{\text{p}} - c_{\text{v}} = VT \frac{\alpha_{\text{p}}^2}{\kappa_{\text{T}}} \quad (12)$$

where V is the molecular volume. We can take advantage of this relation in two ways, first by computing α_{p} and κ_{T} from our simulations and then computing the constant pressure heat capacity based on the constant volume heat capacity. By using experimental data for α_{p} and κ_{T} , we can also establish “experimental” constant volume heat capacities, which are difficult to measure directly. In this work, we have done both, as detailed in the Results and Discussion sections.

The classical—that is, without any quantum corrections—heat capacity $c_{\text{p}}^{\text{class}}$ can be obtained from the fluctuations in the

Table 2. Statistics of a Linear Fit of Calculated to Experimental Values According to $y = ax + b^a$

force field	N	a	b	RMSD	% dev.	R^2
ρ (g/l)						
GAFF	235	0.96	58.5	82.9	4	97%
OPLS/AA	235	0.98	20.9	40.4	2	99%
CGenFF ³⁷	111	1.03	−36.0	26.0	2	99%
OPLS/AA ⁷⁰	9	1.01	−24.0	45.3	4	96%
ΔH_{vap} (kJ/mol)						
GAFF	231	1.07	0.8	10.6	17	83%
OPLS/AA	231	0.96	3.4	6.5	11	89%
CGenFF ³⁷	95	0.94	2.4	4.7	7	84%
γ (10^{-3} N/m)						
GAFF	155	0.75	0.9	8.6	23	70%
OPLS/AA	155	0.97	−5.5	7.3	22	89%
$\epsilon(0)$						
GAFF	163	0.27	0.4	15.7	35	55%
OPLS/AA	176	0.16	0.7	15.9	43	55%
α_{p} (10^{-3} /K)						
GAFF	221	0.90	0.3	0.3	24	67%
OPLS/AA	221	0.91	0.3	0.3	21	75%
OPLS/AA ⁷⁰	9	0.53	0.8	0.7	42	39%
κ_{T} (1/GPa)						
GAFF	103	0.66	0.0	0.3	27	74%
OPLS/AA	103	0.76	0.1	0.3	19	85%
OPLS/AA ⁷⁰	8	0.93	0.0	1.1	59	84%
c_{p} (J/mol K)						
GAFF	130	1.08	−30.9	19.8	10	98%
OPLS/AA	132	1.10	−30.2	18.2	10	97%
OPLS/AA ⁷⁰	9	0.94	3.5	10.4	7	94%
c_{v} (J/mol K)						
GAFF	72	1.02	−17.6	18.8	10	97%
OPLS/AA	72	1.04	−17.9	18.3	9	95%
OPLS/AA ⁷⁰	8	1.01	−5.4	10.8	7	95%
$c_{\text{p}}^{\text{class}}$ (J/mol K)						
GAFF	214	1.77	−21.6	148.3	77	87%
OPLS/AA	214	1.98	−52.8	147.0	73	93%

^aUncertainties in the simulation results are used as weights in the fit. The number of (experimental) data points N is given for each property. Root mean square deviation (RMSD) from experimental values, average relative deviation in percent, and the correlation coefficient R^2 are given. OPLS/AA results from ref 70 and CGenFF results from ref 37 (using the so called CHARMM generalized force field) are also listed for comparison.

enthalpy:³⁸

$$k_{\text{B}}T^2 c_{\text{p}}^{\text{class}} = \langle \delta H^2 \rangle \quad (13)$$

Although this is straightforward to calculate, the numbers obtained in this manner are a factor of 2 too high (Table 2). Therefore, we have determined the heat capacities c_{p} and c_{v} on the basis of the two phase thermodynamic method^{68–70} (described in the Supporting Information), which is based on the convolution of the density of states with a weighting function based on

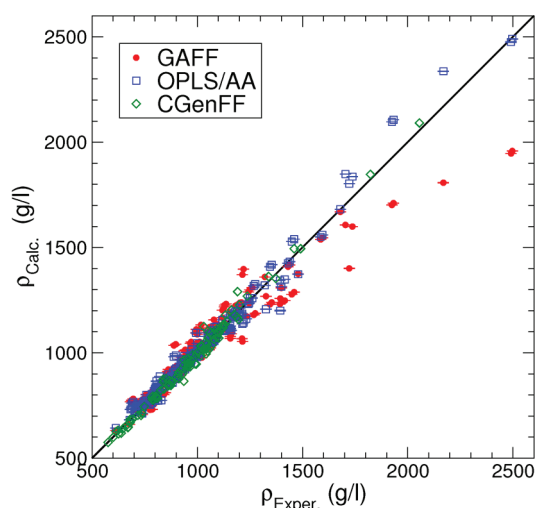


Figure 1. Correlation between densities (ρ) calculated by MD simulation using GAFF, OPLS/AA, CGenFF, and experimental results. The CGenFF data were adopted from Vanommeslaeghe et al.³⁷ and are based on a different (but similar) set of molecules, including 111 molecules. For a full list of the CGenFF data, we refer to the reference and the supplemental files therein.

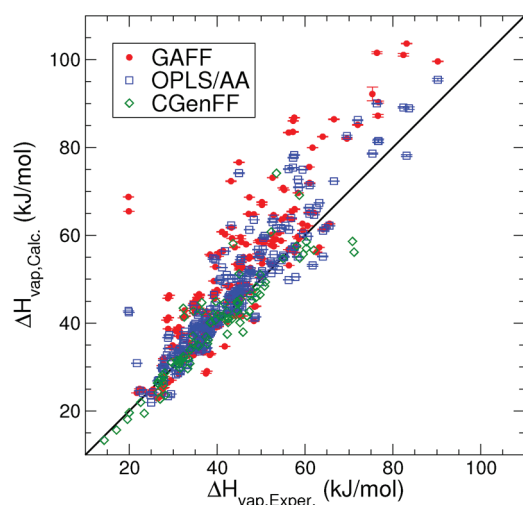


Figure 2. Correlation between enthalpy of vaporization (ΔH_{vap}) calculated using GAFF, OPLS/AA, CGenFF, and experimental results. The CGenFF data were adopted from Vanommeslaeghe et al.³⁷ and are based on a different (but similar) set of molecules, including 95 molecules. For a full list of the CGenFF data, we refer to the reference and the supplemental files therein.

quantum harmonic oscillators, as introduced originally by Berens et al.⁷¹ The final expression yielding the heat capacity c_V is

$$c_V = k_B \int_0^\infty [\text{DoS}_{\text{gas}}(\nu) W_{\text{gas}}^{\text{cv}}(\nu) + \text{DoS}_{\text{solid}}(\nu) W_{\text{solid}}^{\text{cv}}(\nu)] d\nu \quad (14)$$

DoS_{gas} and $\text{DoS}_{\text{solid}}$ denote the density of states in a gas and a solid, $W_{\text{gas}}^{\text{cv}}(\nu)$ and $W_{\text{solid}}^{\text{cv}}(\nu)$ are weighting factors for the same, and c_P can be obtained by combining eq 12 and eq 14. For all details and a complete derivation, we refer the reader to the Supporting Information.

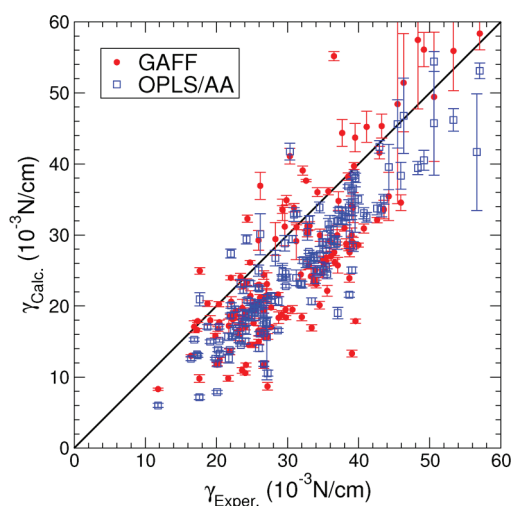


Figure 3. Correlation between surface tension (γ) calculated using the GAFF and the OPLS/AA force fields and experimental results.

The properties investigated fall into two categories: those that follow directly from the ensemble average of a property (energy, pressure, volume) and those based on fluctuations (heat capacities, compressibility, and expansion coefficient). For the first category, error estimates were based on a block averaging procedure that automatically takes the autocorrelation of the property under investigation into account.⁷² Properties like potential energy and density usually have relatively short autocorrelation times. The surface tension fluctuates significantly but also has a short autocorrelation time. For the second category, we have used a different approach when estimating the error. By dividing the entire simulation trajectory into nine, in time, equally long parts, we get nine values for each property, from which we can estimate the total error. In the case of c_V , we used five blocks of 20 ps for error estimation instead.

We calculated c_P on the basis of eq 12 and estimated the error δc_P from the errors in c_V (δc_V), α_P ($\delta \alpha_P$), and κ_T ($\delta \kappa_T$) as

$$\delta c_P^2 = \delta c_V^2 + \left(\frac{2VT\alpha_P}{\kappa_T}\right)^2 \delta \alpha_P^2 + \left(\frac{VT\alpha_P^2}{\kappa_T^2}\right)^2 \delta \kappa_T^2 \quad (15)$$

or, expressed in Δc (eq 12):

$$\delta c_P^2 = \delta c_V^2 + (\Delta c)^2 \left(2\frac{\delta \alpha_P^2}{\alpha_P^2} + \frac{\delta \kappa_T^2}{\kappa_T^2}\right) \quad (16)$$

3. RESULTS

Correlations between experimental data and simulations for observables and derived quantities are plotted in Figures 1–8. The statistics for linear fits to the data ($y_{\text{calcd}} = ay_{\text{exptl}} + b$) are given in Table 2 for each of the observables and the two force fields, plus similar data from refs 37 and 70. To identify which specific molecule generated a certain value in the figures, we refer to Tables S2–S10 in the Supporting Information. An overview of the names of the molecules, their formula, molecular weight, CAS number, and ChemSpider ID is given in Table S1 (Supporting Information). For many molecules, results at different

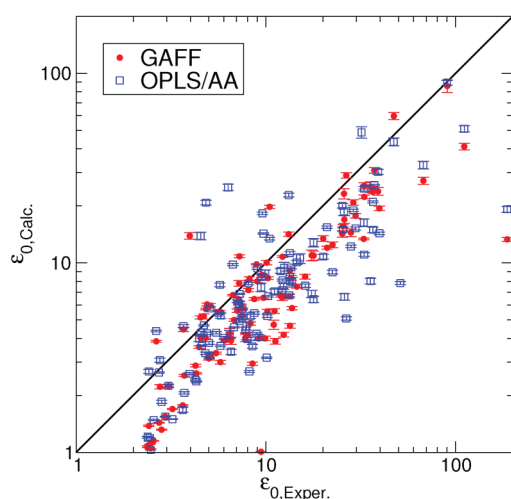


Figure 4. Correlation between dielectric constant (ϵ_0) calculated using the GAFF and the OPLS/AA force fields and experimental results. Note the logarithmic axes.

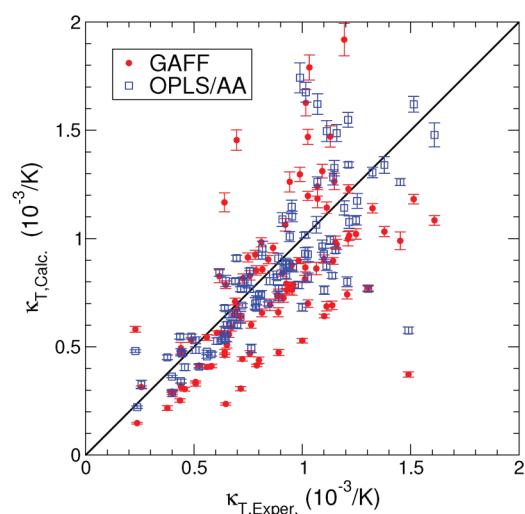


Figure 6. Correlation between isothermal compressibility (κ_T) calculated using the GAFF and the OPLS/AA force fields and experimental results.

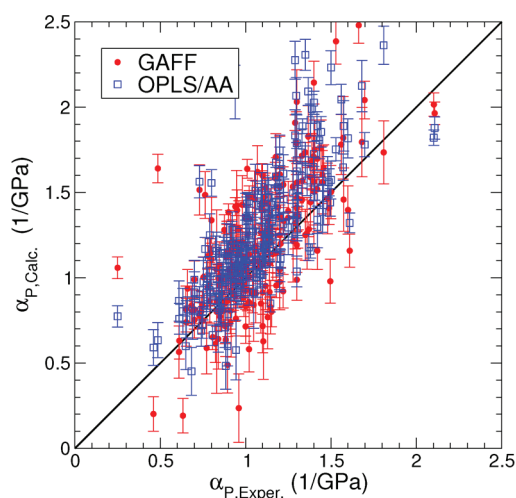


Figure 5. Correlation between volumetric expansion coefficient (α_p) calculated using the GAFF and the OPLS/AA force fields and experimental results.

temperatures were generated, and hence the number of data points may be larger than the number of molecules. For densities, heats of vaporization, surface tensions, and dielectric constants, some of the experimental values were generated from analytical functions of temperature based on experimental data, the parameters of which are given in the Handbook of Chemistry and Physics,⁷³ the Landolt-Bornstein database,⁷⁴ and Yaws' book on Thermophysical Properties of Chemicals and Hydrocarbons.⁷⁵ In addition, we parametrized the dielectric constant, heat capacity at constant pressure, and isothermal compressibility as a function of the temperature for some molecules (see below).

3.1. Statistics. In the following, we discuss general trends in all properties first; outliers are described separately below. A comparison of the values in Table 2 shows that OPLS/AA is slightly better than GAFF at reproducing experimental data for most observables, with both lower RMSD and higher correlation coefficients R^2 .

3.1.1. Density. The density ρ (Figure 1, Table S2) of virtually all liquids is reproduced very well, with $R^2 = 97\%$ (GAFF) and

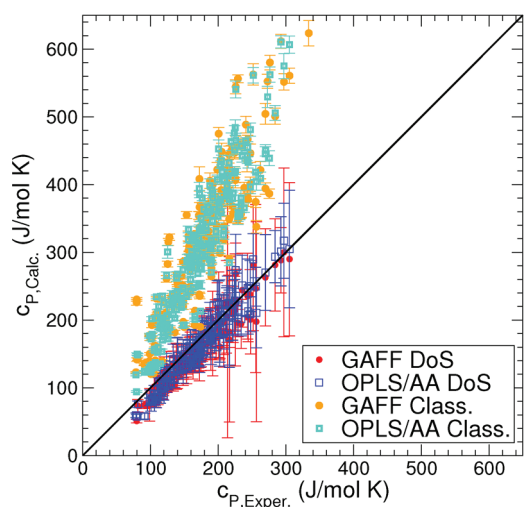


Figure 7. Correlation between measured heat capacity at constant pressure (c_p) and computed using the GAFF and the OPLS/AA force fields based on either the density of states (DoS) method, which includes quantum corrections and a Δc correction based on simulations, or based on a purely classical treatment (c_p^{class} , Class.).

99% (OPLS/AA) (Table 2). For GAFF, the densities are systematically slightly underestimated ($a = 0.96$), while for OPLS/AA, $a = 0.98$, very close to 1, and both have an R^2 close to 100%. In a recent publication, Vanommeslaeghe et al. presented the CHARMM general force field (CGenFF).³⁷ They calculated densities for a set of 111 drug-like molecules, using boxes of 216 molecules. Their reported densities are also very accurate with $a = 1.03$ and $R^2 = 99\%$, see Figure 1 and Table S2.

3.1.2. Enthalpy of Vaporization. ΔH_{vap} (Figure 2, Table S4) correlates very well with experimental data in most cases, with $R^2 = 83\%$ (GAFF) and 89% (OPLS/AA) (Table 2). The GAFF overestimates ΔH_{vap} with slope $a = 1.07$, while OPLS/AA underestimates a slightly at 0.96. These deviations cannot just be attributed to a small number of outliers, as may be evident from Figure 2. Vanommeslaeghe et al.³⁷ calculated enthalpy of

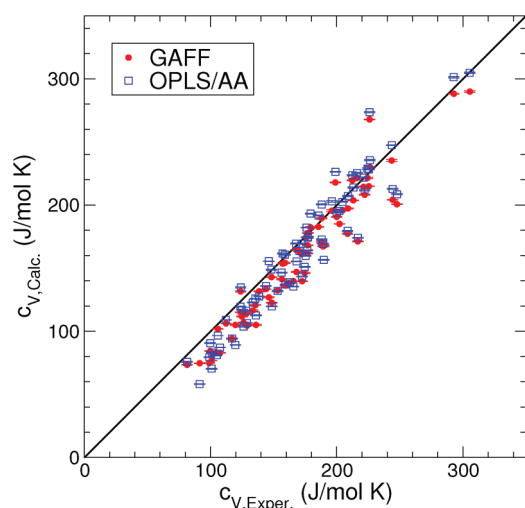


Figure 8. Correlation between measured heat capacity at constant volume (c_v) and computed using the GAFF and the OPLS/AA force fields based on the density of states method, which includes quantum corrections.

vaporization for a set of 95 small molecules. Like for OPLS/AA, ΔH_{vap} is underestimated in CGenFF calculations with a slope of $a = 0.94$. The correlation between experiments and simulation is similar to the two force fields studied here, $R^2 = 84\%$. The CGenFF data set is based on a comparable but different set of molecules than what has been analyzed here (37 molecules overlap between the two studies). To simplify a comparison between OPLS/AA, GAFF, and CGenFF, we have listed the CGenFF ΔH_{vap} values from the study by Vanommeslaeghe et al. next to OPLS/AA and GAFF values in Table S3, and we have plotted them in Figure 2.

3.1.3. Surface Tension. The surface tension γ (Figure 3, Table S4) seems to be underestimated systematically in both force fields with slope $a = 0.75$ (GAFF) and 0.97 (OPLS/AA, Table 2). The interactions between molecules on the surface are not sufficiently strong, a well known problem with nonpolarizable force fields.^{21,25,76} The values are spread around the diagonal for both GAFF ($R^2 = 70\%$) and OPLS/AA ($R^2 = 89\%$), and here again OPLS/AA performs slightly better than GAFF.

3.1.4. Dielectric Constant. For 32 molecules, a novel parametrization of the temperature dependence of the dielectric constant was made on the basis of experimental values predominantly from the Landolt-Bornstein database.⁷⁷ The parametrization is to a polynomial of second or third order (as is used in the Handbook of Chemistry and Physics⁷³), and the resulting coefficients are given in Table 3. Interpolations of these polynomials were used in order to compare the simulations to experimental data, and the fits are presented in Figure S3 of the Supporting Information.

The dielectric constant $\epsilon(0)$ (Figure 4, Table S5) appears to be the most difficult property to reproduce in our simulations, with slopes $a < 0.5$ and $R^2 \leq 60\%$ for both force fields (Table 2). Apart from lacking explicit polarization, limited sampling (1000 molecules for 10 ns were used in all cases) may be one of the causes; another contributing factor is the high viscosity of molecules containing alcohol or amine groups, further aggravated by the fact that some of these molecules were simulated at temperatures close to the melting temperature.

Some liquids have extremely large dielectric constants, e.g., methanamide ($\epsilon(0) = 109$) and *N*-methylformamide ($\epsilon(0) = 190$).

For these molecules, GAFF predicts 41 and 14, respectively, while OPLS/AA predicts 51 and 19. Xie et al. report a simulated dielectric constant of 200 for *N*-methylformamide, using a polarizable model, with only 256 molecules and 1 ns of simulation, but the authors state that “The dielectric constants have only been averaged for 1 ns of simulation time, and they are almost certain not yet converged.” Indeed, Whitfield et al. had previously concluded that very long times (50 ns) may be needed to obtain converged dielectric constants of molecules like *N*-methylacetamide because they tend to form long linear chains.⁷⁸ Such chains can in periodic simulation systems become “infinite”, which may contribute to the long relaxation time. It should be noted, however, that for most molecules in our study, the values are well converged, as witnessed by small error bars. Deviations from experimental results are therefore due predominantly to a lack of polarization and too low mobility of molecules. Interestingly, GAFF is somewhat better at predicting $\epsilon(0)$ than OPLS/AA (Table 2), most likely because the partial charges are somewhat higher for most molecules.

3.1.5. Volumetric Expansion Coefficient. The volumetric expansion coefficient α_p is plotted in Figure 5 and tabulated in Table S6. The slope of the correlation plots is slightly less than 1 for both GAFF ($a = 0.9$) and OPLS/AA ($a = 0.91$), and there is a large spread around the $y = x$ line for both OPLS/AA and GAFF with a RMSD of 0.3/GPa in both cases.

3.1.6. Isothermal Compressibility. For 53 molecules, an interpolation of experimental values of the isothermal compressibility κ_T as a function of the temperature was performed (Table 4 and Figure S3). The simulated κ_T 's are plotted versus the experimental values in Figure 6 and tabulated in Table S8. Like for α_p , the spread in numbers is large, and the slope of the correlation plots is significantly less than 1 (GAFF, 0.66; OPLS/AA, 0.76, Table 2). In general, it seems that fluctuation properties are more difficult to predict than simple linear averages. Although we applied a very strict convergence criterion for the total energy of 0.5 J/mol/ns per degree of freedom, it may be that even longer equilibration times and production simulations are needed.

3.1.7. Heat Capacities. For three molecules, an interpolation of experimental numbers is presented in Table 5 and Figure S4. The heat capacity is a difficult property to calculate due to significant quantum effects. The simple eq 7 produces numbers (c_p^{class}) that are twice too high (Table 2). Since the energy taken up by vibrations in a classical harmonic oscillator is much higher than for a quantum harmonic oscillator at the same frequency, the c_p^{class} values are too high. Introducing quantum corrections, in the manner proposed by Berens et al.,⁷¹ on which the two phase thermodynamics (2PT) method^{68–70} is based, presupposes that the frequencies in the classical simulation are correct: this is often the case since most force constants have been derived from spectroscopic experiments. It should be noted that there is no *a priori* reason to assume that the intermolecular degrees of freedom behave harmonically, as they are determined by Coulomb and van der Waals interactions. Despite these theoretical shortcomings, the 2PT method produces reasonable results for c_p (see Figure 7, Table 2, and Table S8)—much closer to experiment than c_p^{class} on any account. In order to compute c_p , it is necessary to add a correction Δc (eq 12) to the heat capacity at constant volume c_v that is produced by the density of states analysis. Δc is underestimated by classical force field calculations; however, c_p still is estimated reasonably, with $a = 1.08$ for GAFF and $a = 1.02$ for OPLS/AA with correlation coefficients $R^2 = 98\%$ and 97% , respectively. If we compare just c_v from our simulations (i.e., without adding in Δc) and subtract the experimental

Table 3. Parameterization of Temperature Dependence of Dielectric Constants in a Polynomial Form $\epsilon(0) = A + BT + CT^2 + DT^3$, Which Is the Same Form Used in the Handbook of Chemistry and Physics^{73a}

molecule	<i>N</i>	χ^2	T_{\min}	T_{\max}	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
bromomethane	12	0.7	194.60	275.70	52.59	−2.812e−01	4.565e−04	0
methanol	92	1.0	175.62	337.75	226.69	−1.319e+00	2.937e−03	−2.359e−06
1,1,1,2,2-pentachloroethane	9	0.0	245.15	338.15	13.81	−5.527e−02	7.186e−05	0
1,1,2,2-tetrachloroethane	14	0.2	231.15	318.15	71.61	−3.630e−01	5.010e−04	0
1,2-dibromoethane	39	0.1	288.15	353.15	10.31	−3.114e−02	4.200e−05	0
1,1-dichloroethane	8	0.2	288.15	323.15	36.77	−1.300e−01	1.361e−04	0
2-chloroethanol	30	3.1	263.15	401.75	105.36	−3.245e−01	3.619e−05	5.019e−07
ethanamide	7	0.3	358.15	448.20	−200.55	1.551e+00	−2.239e−03	0
methylsulfonylmethane	6	0.0	293.15	323.15	53.55	−2.539e−01	3.571e−04	0
2-aminoethanol	7	0.4	283.65	298.15	166.68	−7.576e−01	1.018e−03	0
1,3-dioxolan-2-one	24	0.5	309.46	364.15	223.34	−4.560e−01	9.143e−05	0
1,3-dioxolane	31	0.2	175.93	303.15	40.61	−2.507e−01	6.323e−04	−5.695e−07
dimethoxymethane	5	0.0	170.65	298.15	2.59	−9.298e−04	3.847e−06	0
ethylsulfonylethane	6	0.1	293.15	323.15	11.68	−1.994e−02	0.000e+00	0
2-methylpropan-2-amine	4	0.0	291.15	303.15	294.70	−1.887e+00	3.060e−03	0
thiophene	14	0.1	252.65	333.15	2.32	5.071e−03	−1.232e−05	0
furan	31	0.2	198.15	303.15	6.69	−2.044e−02	2.644e−05	0
pentane-2,4-dione	9	2.0	291.15	323.15	−532.57	3.658e+00	−5.982e−03	0
3-methylpyridine	6	1.0	293.15	333.00	35.54	−9.303e−02	4.307e−05	0
benzenethiol	6	0.3	293.15	358.15	5.72	−7.033e−03	7.362e−06	0
(E)-hex-2-ene	6	0.0	157.00	295.00	2.43	−1.132e−03	−1.372e−06	0
1-methoxy-2-(2-methoxyethoxy)ethane	5	0.0	298.15	333.15	32.07	−1.359e−01	1.766e−04	0
diethyl propanedioate	7	0.2	293.15	343.15	19.98	−5.034e−02	3.345e−05	0
2,4,6-trimethylpyridine	10	0.1	293.15	358.15	16.67	−3.036e−02	2.361e−06	0
triethyl phosphate	6	0.1	294.15	333.15	−1.59	1.317e−01	−2.780e−04	0
phenylmethanol	26	0.2	278.15	363.15	105.48	−5.130e−01	6.802e−04	0
tetrahydrothiophene 1,1-dioxide	57	0.4	300.75	398.15	488.81	−3.732e+00	1.055e−02	−1.017e−05
2,4,6-trimethylpyridine	10	0.1	293.15	358.15	16.67	−3.036e−02	2.361e−06	0
dimethoxymethane	5	0.0	170.65	298.15	2.92	−4.106e−03	1.126e−05	0
1,3-dichloropropane	5	0.1	298.15	333.15	−61.39	4.818e−01	−8.107e−04	0
methylsulfonylmethane	6	0.2	273.30	310.48	23.41	−8.896e−02	1.076e−04	0
1,2-ethanedithiol	3	0.0	293.15	333.15	11.23	−1.350e−02	0.000e+00	0

^a T_{\min} and T_{\max} (K) indicate the validity range of the parameterization. *N* indicates the number of points in the fit; χ^2 is the root mean square deviation. See the Supporting Information for details.

Δc from the measured c_p , we find a very good correlation (GAFF, $a = 1.02$, $R^2 = 97\%$; OPLS/AA, $a = 1.04$, $R^2 = 95\%$), see Figure 8 and Table 2. Although correlation between experimental results and calculations can by no means validate the underlying theoretical model, it nevertheless indicates that the results are meaningful, because we have approximately 70 experimental c_V values to which to compare. Indeed, although the root-mean-square deviation (RMSD) from experimental results is similar for c_V and c_p , the fit to experimental results is much better (slope a close to 1) for both OPLS/AA and GAFF. The DOS simulations were performed without constraints, and the heat capacities depend directly on the intra- and intermolecular vibrations. Deviations from the experimental heat capacities could therefore indicate problems with the force constants for intramolecular motions.

3.2. Outliers Per Force Field. Table 6 shows how the molecular models of the individual molecules perform relative to the force field as a whole. The average relative deviations in σ and averaged over 1–8 data points (depending on the availability of experimental data) signals how well the force field performs for each molecule. The properties used were density, enthalpy of

vaporization, surface tension, dielectric constant, volumetric expansion coefficient, isothermal compressibility, and the heat capacity at constant volume.

Some types of molecules are problematic in both of the force fields considered here. Small molecules containing more than one Cl or Br atom generally have both density and enthalpy of vaporization values that deviate significantly from experimental reference. This is not the case for molecules containing only one of these atoms, or molecules where there is a spacer (e.g., a CH_2 group) between them. It could therefore be that the differences are caused by overlapping atoms. By introducing a new atom type of Br and Cl for the case where there are two such atoms next to each other on the carbon chain, these problems might be resolved.

The density and enthalpy of vaporization of methanoic acid (formic acid) were particularly hard to reproduce, as was noted previously by Jedlovsky and Turi, who constructed a specific potential for this molecule.⁷⁹ The main feature responsible for the improved model in this case was a higher charge ($\approx 0.1e$) on the C–H atom than is used in either OPLS/AA (0) or GAFF (0.04). Methanoic acid forms very strong linear chains, which are

Table 4. Parameterization of Temperature Dependence of Isothermal Compressibility Constants in a Polynomial Form $\kappa_T = A + BT + CT^{2a}$

molecule	<i>N</i>	χ^2	T_{\min}	T_{\max}	<i>A</i>	<i>B</i>	<i>C</i>
dichloromethane	3	0.000	293.15	303.15	-1.709e+01	1.144e-01	-1.800e-04
methanamide	5	0.008	288.15	323.15	1.352e-01	9.161e-04	0
nitromethane	4	0.020	298.15	323.15	-1.253e+00	6.666e-03	0
methanol	24	0.014	213.15	333.15	1.004e+00	-6.791e-03	2.557e-05
acetonitrile	5	0.000	298.15	318.15	3.174e+00	-2.209e-02	5.114e-05
1,1,2,2-tetrachloroethane	2	0.000	293.15	303.15	-4.962e-01	3.900e-03	0
1,1,2-trichloroethane	7	0.002	288.15	318.15	-7.213e-01	4.937e-03	0
bromoethane	5	0.010	273.15	323.15	9.748e+00	-6.685e-02	1.287e-04
N-methylformamide	4	0.011	288.15	313.15	6.378e-03	1.968e-03	0
nitroethane	3	0.015	298.15	323.15	-9.873e-01	6.004e-03	0
ethanol	16	0.007	203.15	363.15	1.280e+00	-8.946e-03	2.857e-05
methylsulfinylmethane	7	0.030	293.15	353.15	5.206e-01	-3.136e-03	1.052e-05
2-aminoethanol	6	0.000	278.15	333.15	7.273e-01	-4.276e-03	1.051e-05
1,3-dichloropropane	6	0.000	283.15	323.15	6.932e-01	-4.785e-03	1.678e-05
propan-2-one	10	0.010	293.15	328.15	-3.053e+00	1.468e-02	0
methyl acetate	8	0.012	293.15	328.15	-2.562e+00	1.249e-02	0
1,3-dioxolane	2	0.000	293.15	313.15	-1.317e+00	6.960e-03	0
1-bromopropane	7	0.003	288.15	318.15	-1.264e+00	8.037e-03	0
N,N-dimethylformamide	18	0.018	288.15	333.20	1.748e+00	-1.073e-02	2.367e-05
1-nitropropane	3	0.004	298.15	323.15	-1.111e+00	6.420e-03	0
2-nitropropane	3	0.020	298.15	323.15	-1.060e+00	6.604e-03	0
1,4-dichlorobutane	5	0.004	288.15	318.15	-8.725e-01	5.246e-03	0
propane-1,2,3-triol	19	0.003	293.15	473.15	8.358e-01	-4.323e-03	7.862e-06
propan-1-amine	6	0.036	293.15	323.15	-2.469e+00	1.238e-02	0
N,N-dimethylacetamide	5	0.015	288.15	318.15	-5.890e-01	4.142e-03	0
butan-1-ol	15	0.021	293.15	393.15	1.307e+00	-8.833e-03	2.543e-05
N-ethylethanamine	5	0.002	298.15	318.15	7.548e+00	-5.536e-02	1.188e-04
butan-1-amine	8	0.003	298.15	328.15	2.330e+00	-1.702e-02	4.371e-05
ethyl acetate	9	0.012	298.15	350.30	5.084e+00	-3.567e-02	7.598e-05
oxolane	5	0.001	278.15	323.15	-9.434e-01	4.999e-03	4.886e-06
1-bromobutane	12	0.000	298.15	333.15	2.650e+00	-1.860e-02	4.413e-05
1-chlorobutane	10	0.029	293.15	318.15	-2.399e+00	1.205e-02	0
pentanenitrile	5	0.005	283.15	323.15	8.811e-01	-7.004e-03	2.429e-05
ethyl propanoate	15	0.022	278.15	338.15	6.964e-01	-7.128e-03	2.882e-05
2-methylbutan-2-ol	2	0.000	293.15	298.15	-1.495e+00	8.600e-03	0
pentan-1-ol	8	0.010	293.15	333.15	3.158e+00	-2.044e-02	4.292e-05
pentan-3-ol	10	0.003	293.15	368.15	4.952e+00	-3.315e-02	6.587e-05
nitrobenzene	5	0.009	298.15	323.15	-3.337e-01	2.832e-03	0
cyclohexanone	5	0.021	298.15	308.15	-9.399e-01	5.421e-03	0
hexan-2-one	8	0.022	278.15	338.15	-1.451e+00	8.315e-03	0
1-methoxy-2-(2-methoxyethoxy)ethane	6	0.001	298.15	318.15	-8.794e-01	5.105e-03	0
N,N-diethylethanamine	8	0.006	298.15	328.15	4.400e+00	-3.405e-02	8.064e-05
N-propan-2-ylpropan-2-amine	7	0.001	298.15	328.15	9.459e+00	-6.732e-02	1.357e-04
methoxybenzene	5	0.043	298.15	338.15	-1.520e+00	7.287e-03	0
3-methylphenol	6	0.041	298.15	413.15	1.744e+00	-1.029e-02	2.104e-05
toluene	50	0.006	288.15	333.15	2.342e+00	-1.627e-02	3.853e-05
diethyl propanedioate	7	0.000	298.15	328.15	2.164e+00	-1.397e-02	3.048e-05
heptan-2-one	2	0.000	293.15	298.15	-8.915e-01	6.200e-03	0
ethylbenzene	7	0.008	293.15	333.15	2.524e+00	-1.652e-02	3.683e-05
1,2-dimethylbenzene	10	0.022	273.15	417.50	-2.914e-01	1.846e-03	6.429e-06
octan-1-ol	16	0.033	293.15	413.15	2.242e+00	-1.449e-02	3.206e-05
quinoline	2	0.000	333.15	373.15	-5.477e-01	3.320e-03	0
(1-methylethyl)benzene	3	0.003	293.15	298.15	-6.340e-01	5.110e-03	0

^a T_{\min} and T_{\max} (K) indicate the validity range of the parameterization. *N* indicates the number of points in the fit; χ^2 is the root mean square deviation. See the Supporting Information for details.

Table 5. Parameterization of Temperature Dependence of Heat Capacity at Constant Pressure in a Polynomial Form $c_p = A + BT^a$

molecule	N	χ^2	T_{\min}	T_{\max}	A	B
1,3-dioxolane	9	0.187	288.15	328.15	4.371e+01	2.613e-01
1,2,3,4-tetrafluorobenzene	41	0.145	235.47	319.79	1.158e+02	2.491e-01
1,2,3,5-tetrafluorobenzene	25	0.343	229.32	311.18	1.186e+02	2.400e-01

^a T_{\min} and T_{\max} (K) indicate the validity range of the parameterization. N indicates the number of points in the fit; χ^2 is the root mean square deviation. See the Supporting Information for details.

difficult to break. This leads to long correlation times for the system dipoles and to dielectric constants that are far from the experimental values (Table S5).

Benzaldehyde and furan are also problematic in both force fields. Even if they both generate decent densities and enthalpies of vaporization, the other properties (surface tension, dielectric constant, and thermal expansion coefficient) are far from the experimental values.

Molecules containing a nitro group (specially nitromethane, 1-nitropropane, and 2-nitropropane) stick out as a problematic group in GAFF. The charges on nitro groups are high, leading to high density and enthalpy of vaporization.

The standard OPLS/AA parameterization of alcohols has been reported to perform poorly for octan-1-ol. MacCallum and Tieleman⁸⁰ therefore derived a specific united atom potential of the molecule where they used modified charges on the headgroup. The OPLS/AA parametrization investigated here gives both too high a density and too high an enthalpy of vaporization, and therefore the other properties investigated for this molecules also deviate from experimental results. Methyl-2-methylprop-2-enoate shows similar problems, and this could probably be corrected in a similar way. It should be noted that, compared to GAFF, the charges on the headgroup in these two molecules are relatively high in OPLS/AA.

4. DISCUSSION

The development of force fields for molecular simulation is critically dependent on the availability of good reference data, preferably from experimental sources. All force fields, be they empirical, purely derived from quantum-mechanics, or a combination of the two, will eventually have to face the test of comparing predicted to measured values. There is a large amount of literature on force field testing for proteins and peptides,^{57,81–87} nucleic acids,^{88–91} carbohydrates,⁹² specific organic molecules or protein fragments,^{20,26,93–97} and ions,^{98–101} to list but a few. In addition, there are indirect force field tests, for instance of the binding energy in protein–ligand complexes,^{16,102} protein structure prediction,¹⁰³ or of force-field-based docking codes.^{104–106} It is interesting to mention the industrial fluid properties simulation challenges, which are stimulating modelers to predict properties of liquids by any means, including molecular simulation.^{107,108}

Here, we have introduced a benchmark set of 146 liquids in order to assess two popular all atom force fields, OPLS/AA and GAFF, and to set a standard for future force fields. For comparison, we have included an independent density and enthalpy of vaporization data set computed using CGenFF, based on a similar set of molecules.³⁷ Calculated density, enthalpy of vaporization, heat capacities, surface tension, dielectric constants,

Table 6. Average Relative Deviation (σ) from Experimental Values, in Brackets, the Number of Observables^a

name	CGenFF	GAFF	OPLS/AA
1. chloroform		2.1(6)	3.0(7)
2. dichloro(fluoro)methane		1.0(4)	1.3(4)
3. dibromomethane		2.9(6)	1.7(7)
4. dichloromethane		1.7(7)	3.6(7)
5. methanal		0.3(4)	0.3(4)
6. methanoic acid		4.5(6)	2.6(7)
7. bromomethane		1.4(3)	0.4(3)
8. methanamide	0.0(1)	1.2(7)	0.4(6)
9. nitromethane		2.0(7)	0.8(7)
10. methanol	0.0(2)	0.8(7)	0.8(7)
11. 1,1,1,2-pentachloroethane		0.5(4)	0.8(4)
12. 1,1,2,2-tetrachloroethane		1.7(7)	1.7(7)
13. 1,1-dichloroethene		1.7(4)	0.8(4)
14. 1,1,2-trichloroethane		1.2(7)	0.9(7)
15. acetonitrile	0.0(1)	1.1(7)	2.2(7)
16. 1,2-dibromoethane		2.6(7)	4.0(7)
17. 1,1-dichloroethane	0.0(1)	0.7(7)	1.7(7)
18. 1,2-dichloroethane		1.6(7)	1.2(7)
19. methyl formate		0.9(4)	0.8(5)
20. bromoethane	0.0(1)	2.2(7)	0.6(7)
21. chloroethane	0.0(1)	0.8(5)	1.3(5)
22. 2-chloroethanol		0.4(4)	0.5(4)
23. ethanamide		0.2(4)	0.8(5)
24. N-methylformamide		1.4(7)	1.4(7)
25. nitroethane		1.5(7)	0.7(7)
26. methoxymethane		0.5(5)	1.3(5)
27. ethanol	0.0(2)	1.0(7)	0.7(6)
28. 1,2-ethanedithiol		0.6(3)	0.1(3)
29. methylsulfanylmethane	0.1(2)	1.2(5)	1.6(5)
30. methylsulfanylmethane	0.1(1)	1.0(7)	0.6(7)
31. methylsulfanylmethane		1.4(5)	1.2(5)
32. 2-aminoethanol		1.2(5)	1.3(6)
33. ethane-1,2-diamine		1.2(7)	1.9(7)
34. prop-2-enenitrile		1.0(5)	1.2(5)
35. 1,3-dioxolan-2-one		0.5(5)	0.2(4)
36. propanenitrile		1.1(7)	1.9(7)
37. 1,2-dibromopropane		1.1(5)	0.6(4)
38. 1,3-dichloropropane		0.9(7)	1.0(7)
39. (2R)-2-methyloxirane		0.0(2)	0.1(2)
40. propan-2-one	0.0(2)	1.0(7)	0.7(7)
41. methyl acetate	0.0(2)	1.3(7)	0.9(7)
42. 1,3-dioxolane	0.0(1)	1.2(4)	0.6(4)
43. 2-iodopropane		0.7(5)	1.1(5)
44. 1-bromopropane		1.3(7)	0.6(7)
45. N,N-dimethylformamide		0.7(6)	0.5(6)
46. N-methylacetamide	0.0(1)	0.4(4)	0.2(4)
47. 1-nitropropane		1.6(7)	1.2(7)
48. 2-nitropropane		1.6(7)	0.9(7)
49. dimethoxymethane		0.8(5)	0.9(5)
50. propane-1,2,3-triol		1.3(6)	0.8(6)
51. propan-1-amine		1.1(7)	1.5(7)
52. propan-2-amine		0.7(5)	0.6(4)
53. 2-methylpropane	0.0(1)	0.8(5)	1.1(5)

Table 6. Continued

name	CGenFF	GAFF	OPLS/AA
54. ethylsulfanylethane		0.6(5)	0.7(5)
55. butane-1-thiol		0.9(5)	0.5(5)
56. butan-1-ol		1.1(7)	0.9(7)
57. 2-methylpropan-2-ol		0.4(2)	0.1(2)
58. butane-1,4-diol		0.9(6)	0.4(6)
59. (2-hydroxyethoxy)ethan-2-ol		1.2(4)	1.1(5)
60. N-ethylethanamine		1.1(7)	1.2(7)
61. butan-1-amine		1.1(7)	0.9(7)
62. 2-methylpropan-2-amine		1.0(5)	0.8(5)
63. 2-(2-hydroxyethylamino)ethanol		0.5(4)	0.4(4)
64. pyrimidine	0.0(2)	0.7(4)	0.6(4)
65. furan	0.2(2)	1.9(5)	1.9(5)
66. thiophene	0.0(2)	0.7(4)	0.3(5)
67. 1H-pyrrole	0.1(1)	1.3(7)	1.1(7)
68. ethenyl acetate		0.5(4)	0.8(4)
69. oxolan-2-one		0.3(3)	0.3(4)
70. acetyl acetate		1.2(4)	1.2(4)
71. 1,4-dichlorobutane		0.6(7)	0.8(7)
72. oxolane		0.6(6)	1.3(7)
73. ethoxyethene		0.3(3)	0.2(3)
74. ethyl acetate	0.0(2)	1.2(7)	1.1(7)
75. tetrahydrothiophene 1,1-dioxide		0.8(4)	0.9(4)
76. thiolane		0.5(4)	0.4(4)
77. 1-bromobutane		1.1(7)	0.7(7)
78. 1-chlorobutane		1.4(7)	1.8(7)
79. pyrrolidine	0.1(1)	1.3(7)	1.3(7)
80. N,N-dimethylacetamide		1.0(7)	0.9(7)
81. morpholine		0.8(5)	0.9(5)
82. pyridine	0.1(2)	0.6(6)	0.9(7)
83. cyclopentanone		0.8(5)	0.6(5)
84. 1-cyclopropylethanone		0.2(2)	0.1(2)
85. pentane-2,4-dione		0.9(5)	1.3(5)
86. methyl 2-methylprop-2-enoate		0.8(5)	3.6(5)
87. pentanenitrile		0.6(6)	1.6(7)
88. ethyl propanoate		1.3(7)	1.1(7)
89. diethyl carbonate		2.1(7)	0.7(6)
90. pentan-1-ol		1.0(7)	0.9(7)
91. pentan-3-ol		1.0(7)	1.1(7)
92. 2-methylbutan-2-ol		1.1(5)	0.5(5)
93. pentane-1,5-diol		0.8(6)	0.6(6)
94. pentan-3-amine		0.5(4)	0.6(4)
95. 1,2,3,4-tetrafluorobenzene		0.2(2)	0.1(2)
96. 1,2,3,5-tetrafluorobenzene		0.2(2)	0.1(2)
97. 1,3-difluorobenzene	0.2(2)	0.7(4)	1.3(5)
98. 1,2-difluorobenzene		0.7(4)	1.0(5)
99. fluorobenzene	0.1(2)	1.6(7)	0.5(6)
100. nitrobenzene	0.0(2)	1.1(7)	1.1(7)
101. 2-chloroaniline		0.9(4)	0.6(4)
102. phenol		0.8(4)	0.9(5)
103. benzenethiol		1.4(5)	1.3(5)
104. 2-methylpyridine		0.3(4)	0.9(5)
105. 3-methylpyridine	0.1(2)	0.8(5)	0.6(5)
106. 4-methylpyridine	0.0(2)	1.1(7)	0.4(6)
107. cyclohexanone		1.0(7)	0.9(7)
108. (E)-hex-2-ene	0.0(2)	0.0(2)	0.0(2)

Table 6. Continued

name	CGenFF	GAFF	OPLS/AA
109. hexan-2-one		0.8(6)	0.9(7)
110. 2,4,6-trimethyl-1,3,5-trioxane		1.6(4)	1.0(4)
111. cyclohexanamine		0.8(5)	0.7(5)
112. 2-propan-2-yloxypropane		3.3(7)	0.9(7)
113. 1-methoxy-2-(2-methoxyethoxy)ethane		1.5(7)	1.1(7)
114. triethyl phosphate		2.8(6)	2.2(6)
115. N,N-diethylethanamine		1.2(7)	1.0(7)
116. N-propan-2-ylpropan-2-amine		0.8(6)	0.6(6)
117. trifluoromethylbenzene		0.8(5)	0.5(4)
118. benzonitrile		1.0(5)	1.0(5)
119. benzaldehyde	0.2(2)	5.7(7)	3.7(6)
120. toluene	0.1(2)	1.6(7)	1.3(7)
121. methoxybenzene	0.1(2)	1.2(7)	1.1(7)
122. phenylmethanol		1.0(5)	0.8(5)
123. 2-methylphenol		0.9(5)	0.8(5)
124. 3-methylphenol		1.0(5)	0.9(5)
125. 4-methylphenol	0.1(1)	1.2(5)	0.7(5)
126. diethyl propanedioate		1.1(4)	0.8(4)
127. 2,4-dimethylpentan-3-one		0.6(4)	0.4(4)
128. heptan-2-one		1.1(7)	0.7(7)
129. ethenylbenzene		1.2(5)	1.1(5)
130. 1-phenylethanone		1.0(7)	1.1(7)
131. methyl benzoate		0.9(7)	1.0(7)
132. methyl 2-hydroxybenzoate		1.1(5)	0.4(4)
133. ethylbenzene	0.1(2)	1.4(7)	1.1(7)
134. 1,2-dimethylbenzene	0.1(1)	1.7(7)	1.0(7)
135. 1,2-dimethoxybenzene		0.4(4)	0.6(5)
136. 2,4,6-trimethylpyridine		0.9(5)	1.0(5)
137. octan-1-ol		0.8(6)	1.7(7)
138. 1-butoxybutane		0.7(4)	1.0(5)
139. N-butylbutan-1-amine		0.9(7)	0.8(7)
140. isoquinoline	0.0(1)	0.7(4)	1.3(4)
141. quinoline	0.1(2)	1.1(7)	1.2(7)
142. (1-methylethyl)benzene	0.1(2)	1.0(6)	0.7(6)
143. 1,2,4-trimethylbenzene	0.1(1)	1.2(6)	1.0(6)
144. 2,6-dimethylheptan-4-one		1.0(5)	0.9(5)
145. 1-chloronaphthalene		0.5(6)	1.3(7)
146. phenoxybenzene		0.6(4)	1.1(5)

^a Average relative deviation larger than 1 σ is printed in bold, larger than 1.5 σ in bold italic.

volumetric expansion coefficients, and isothermal compressibility from the two force fields are compared to experimental values. Indeed the benchmark is quite revealing, in that systematic deviations can be found and rationalized. The knowledge about such deviations will hopefully be useful for further development of the force fields.

To a first approximation, molecular vibrations can be described as quantum harmonic oscillators.¹⁰⁹ Classical harmonic oscillators do not describe the properties of quantum harmonic oscillators, which makes it necessary to implement corrections in computing for instance heat capacities. The two phase thermodynamics method employed here for estimating c_p and c_v relies on the force constants of the force field used, and on the effective frequencies in the simulations. The density of states obtained

from the velocity autocorrelation is convoluted by a weighting function derived from the partition function for a quantum harmonic oscillator in order to obtain a heat capacity for a corresponding quantum liquid. If a force field would allow one to directly reproduce the “correct” density of states, one could use the much simpler fluctuation formulas, as described by Allen and Tildesley;³⁸ however, heat capacities computed in this manner overestimate the experimental values by about 100% for OPLS/AA and GAFF (Table S10). Going beyond the harmonic approximation should therefore be considered by force field developers. Despite efforts in the context of the MMF94 force field¹¹⁰ and the MM3-MM4 family of force fields,^{111–113} this has not been widely adopted in the biomolecular simulation community, although the polarizable AMOEBA force field¹¹⁴ does feature anharmonic bond and angle potentials as well. In principle, it should be advantageous to use for instance Car–Parrinello molecular dynamics,¹¹⁵ in order to more faithfully represent a liquid than is possible in a classical simulation. This was attempted by Kuo et al. for water.¹¹⁶ They find a large scatter in c_p values due to limited sampling, but also a systematic deviation from the experimental value. Obviously, the computational bottleneck that would be introduced by CPMD or related methods will remain difficult to surmount for the immediate future, and therefore force-field-based methods remain necessary. Nevertheless, it is encouraging that there is a trend to use molecular dynamics simulations based on density functional theory codes to study vibrational properties of biomolecular systems beyond the harmonic approximation.^{117–119}

The dielectric constant seems to be the hardest nut to crack. Nonpolarizable force fields (such as GAFF and OPLS/AA) are known to have difficulties in reproducing the dielectric constant and to some extent also the surface tension. In the case of water, for which a large number of force fields have been developed, there are several studies that describe this (for a review, see, for example, Guillot²⁵). Improving the dielectric function often turns out to be done at the cost of the enthalpy of vaporization and the free energy of solvation—properties that may be more important to reproduce in biomolecular simulations. In addition to systematic problems, like sampling or the lack of polarization in our simulations,¹²⁰ the temperature dependence of the dielectric constant provides both a challenge and an opportunity for future force field development. For most molecules, the temperature dependence is very strong, because molecular motion is the largest factor contributing to $\epsilon(0)$. In his review of water models, Guillot has pointed out that the relation between dielectric constant and other properties is complex, and hence it can be used to test and validate force fields, but not likely as a target for force field optimization.²⁵

The benchmark we present here allows one to pinpoint systematic errors in force fields due to the fact that most chemical moieties are represented more than once. The overall performance of GAFF is surprisingly good, seeing that the parameter development was not aimed at liquids. The results from the OPLS/AA force field are slightly better than GAFF, obviously due to the fact that OPLS/AA was parametrized for liquids. The CHARMM generalized force field seems to be even slightly better, at least for density and enthalpy of vaporization.³⁷ It is reassuring for applications of force field calculations beyond liquids that the parameters in most cases are reasonable; however, the results presented here also show that blind faith in force fields is not warranted in all cases. In Table 2, we list the root-mean-square deviation, as well as the average relative deviation,

of the calculated values from the experimental, for each property we have analyzed. Even if our set of molecules is limited to 146, these numbers give a measurement of how well the properties are reproduced in the two force fields, at least for molecules similar to the set presented here.

Wang and Tingjun have recently reported a similar force field test of 71 organic molecules based on the GAFF and OPLS/AA force fields.³² They report densities and enthalpies of vaporization for these molecules and find small deviations from experimental results that are comparable to our numbers. It is encouraging to note that these authors were able to improve the correspondence to experimental numbers by tuning the Lennard-Jones parameters of some of the atom types. How this affects the other properties that we have studied here, in particular, the dielectric constant and the surface tension, remains to be determined, but it is likely that just tweaking the Lennard-Jones parameters is not sufficient to cure the significant and systematic deviations observed for those properties.

Mobley et al. have performed free energy of solvation (ΔG_{hyd}) benchmarks, reporting a RMS error from experimental numbers of 5.2 kJ/mol for more than 500 molecules.^{121,122} This number is comparable to the RMSD of 6.5 kJ/mol we computed for ΔH_{vap} for OPLS/AA (10.6 kJ/mol for GAFF and 4.7 kJ/mol for CGenFF³⁷). Since both numbers are to a large extent determined by the intermolecular energies, we can conclude that the RMS error in intermolecular energies for (“small”) organic molecules is 5–6 kJ/mol using state of the art simulations and nonpolarizable force fields. It should be noted that this result may be biased by the choice of test set, as has been shown in the context of the SAMPL contest where hydration energies were to be predicted.^{123,124} It was found here that larger molecules with multiple functional groups have similar deviations from the experimental hydration energy—errors up to 10 kJ/mol.¹²³ It seems plausible that part of this error is due to the simple nonpolarizable water model used, however, since the enthalpy of vaporization is approximately additive (which can be seen by plotting ΔH_{vap} for, e.g., alkanes as a function of the number of carbons), the error per functional group should still be relatively low, less than 5 kJ/mol for most groups. In the present work, we studied pure liquids only, providing a simpler test set than what has been used in previous studies. Further tests on pure liquids and liquid mixtures should provide a more detailed understanding of the predictive power of force field calculations. At the same time, systematic methods for force field development^{23,125} could be used for the improvement of classical force fields.

■ ASSOCIATED CONTENT

Supporting Information. Complete molecular topologies and structures for use with the GROMACS software suite as well as equilibrated liquid boxes containing coordinates for all 146 systems are available from our Web site <http://virtualchemistry.org>. Simulation parameter files are available in a zip file. The PDF file contains a derivation of the two phase thermodynamics method, as well as four supporting figures and 13 tables. Figure S1 shows ΔD eq 4; Figure S2, the fits to experimental data as a function of temperature for the dielectric constants; Figure S3, the fits to experimental data as a function of temperature for the heat capacity; and Figure S4, the fits to experimental data as a function of temperature for the isothermal compressibility. Tables S11–S13 give the experimental references corresponding to Figures S2–S4 for each molecule. Table S1 contains a list of all

molecules with formula, molecular weight, CAS number, and ChemSpider ID. Full lists of the calculated values for all properties as well as experimental and CGenFF³⁷ reference data (where applicable) are presented for liquid densities (Table S2), enthalpy of vaporization (Table S3), surface tension (Table S4), dielectric constant (Table S5), volumetric expansion coefficients (Table S6), isothermal compressibility (Table S7), heat capacity c_P (Table S8), heat capacity c_V (Table S9), and heat capacity c_P^{class} (Table S10). The tables are presented using the Hill system. This information is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: spoel@xray.bmc.uu.se.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

The Swedish research council is acknowledged for some financial support to D.v.d.S. and for a grant of computer time (SNIC-022-09/10) through the national supercomputer center (NSC) in Linköping and the parallel computing center (PDC) in Stockholm, Sweden. C.C. acknowledges financial support from the Helmholtz Association through the Center for Free-Electron Laser Science. This study was also supported by a Marie Curie Intra-European Fellowship within the seventh European Community Framework Programme to J.S.H. We would like to thank Göran Wallin for stimulating discussions and Marjolein van der Spoel for help with collecting experimental data.

REFERENCES

- Jorgensen, W. L.; Tirado-Rives, J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6665–6670.
- Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157–1174.
- Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2005**, *25*, 1157–1174.
- Hetyenyi, C.; Paragi, G.; Maran, U.; Timar, Z.; Karelson, M.; Penke, B. *J. Am. Chem. Soc.* **2006**, *128*, 1233–1239.
- Case, D. A.; Cheatham, T. E., III; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668–1688.
- Fujitani, H.; Tanida, Y.; Matsuura, A. *Phys. Rev. E* **2009**, *79*, 021914.
- Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1984**, *5*, 129–145.
- Besler, B. H.; Merz, K. M., Jr.; Kollman, P. A. *J. Comput. Chem.* **1990**, *11*, 431–439.
- Francl, M. M.; Carey, C.; Chirlian, L. E.; Gange, D. M. *J. Comput. Chem.* **1996**, *17*, 367–383.
- Francl, M. M.; Chirlian, L. E. *Rev. Comput. Chem.* **2000**, *14*, 1–31.
- Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269–10280.
- Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. *J. Comput. Chem.* **2000**, *21*, 132–146.
- Jakalian, A.; Jack, D. B.; Bayly, C. I. *J. Comput. Chem.* **2002**, *23*, 1623–1641.
- Mobley, D. L.; Dumont, E.; Chodera, J. D.; Dill, K. A. *J. Phys. Chem. B* **2007**, *111*, 2242–2254.
- Mobley, D. L.; Dumont, E.; Chodera, J. D.; Dill, K. A. *J. Phys. Chem. B* **2011**, *115*, 1329–1332.
- Wallin, G.; Nervall, M.; Carlsson, J.; Åqvist, J. *J. Chem. Theory Comput.* **2009**, *5*, 380–395.
- Storer, J. W.; Giesen, D. J.; Cramer, C. J.; Truhlar, D. G. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 87–110.
- Kaminski, G.; Jorgensen, W. *J. Phys. Chem. B* **1998**, *102*, 1787–1796.
- Chandrasekhar, J.; Shariffskul, S.; Jorgensen, W. *J. Phys. Chem. B* **2002**, *106*, 8078–8085.
- Shirts, M. R.; Pitera, J. W.; Swope, W. C.; Pande, V. S. *J. Chem. Phys.* **2003**, *119*, 5740–5761.
- Wensink, E. J. W.; Hoffmann, A. C.; van Maaren, P. J.; van der Spoel, D. *J. Chem. Phys.* **2003**, *119*, 7308–7317.
- van der Spoel, D.; van Maaren, P. J.; Larsson, P.; Timmeanu, N. *J. Phys. Chem. B* **2006**, *110*, 4393–4398.
- van der Spoel, D.; van Maaren, P. J.; Berendsen, H. J. C. *J. Chem. Phys.* **1998**, *108*, 10220–10230.
- Mark, P.; Nilsson, L. *J. Comput. Chem.* **2002**, *23*, 1211–1219.
- Guillot, B. *J. Mol. Liq.* **2002**, *101*, 219–260.
- Hess, B.; van der Vegt, N. F. A. *J. Phys. Chem. B* **2006**, *110*, 17616–17626.
- Caleman, C.; van der Spoel, D. *J. Chem. Phys.* **2006**, *125*, 154508.
- Caleman, C.; van der Spoel, D. *Phys. Chem. Chem. Phys.* **2007**, *9*, 5105–5111.
- Vega, C.; Abascal, J. L. F. *Phys. Chem. Chem. Phys.* **2011**.
- Kaminski, G.; Duffy, E. M.; Matsui, T.; Jorgensen, W. L. *J. Phys. Chem.* **1994**, *98*, 13077–13082.
- Kaminski, G.; Jorgensen, W. L. *J. Phys. Chem.* **1996**, *100*, 18010–18013.
- Wang, J.; Tingjun, H. *J. Chem. Theory Comput.* **2011**, *7*, 2151–2165.
- Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. *Comput. Phys. Commun.* **1995**, *91*, 43–56.
- Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Model.* **2001**, *7*, 306–317.
- van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701–1718.
- Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.
- Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D. *J. Comput. Chem.* **2010**, *31*, 671–690.
- Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford Science Publications: Oxford, 1987.
- Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.
- Schuettelkopf, A. W.; van Aalten, D. M. F. *Acta Crystallogr., Sect. D* **2004**, *60*, 1355–1363.
- Schaftenaar, G.; Noordik, J. H. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 123–134.
- Frisch, M. J.; et al. *Gaussian 03*, Revision C.02; Gaussian, Inc.: Wallingford, CT, 2004.
- Krishnan, R.; Binkley, J.; Seeger, R.; Pople, J. *J. Chem. Phys.* **1980**, *72*, 650–654.
- McLean, A.; Chandler, G. *J. Chem. Phys.* **1980**, *72*, 5639–5648.
- Glukhovtsev, M.; Pross, A.; McGrath, M.; Radom, L. *J. Chem. Phys.* **1995**, *103*, 1878–1885.
- Curtiss, L.; McGrath, M.; Blaudeau, J.; Davis, N.; Binning, R.; Radom, L. *J. Chem. Phys.* **1995**, *103*, 6104–6113.
- Blaudeau, J.; McGrath, M.; Curtiss, L.; Radom, L. *J. Chem. Phys.* **1997**, *107*, 5016–5021.
- Feenstra, K. A.; Hess, B.; Berendsen, H. J. C. *J. Comput. Chem.* **1999**, *20*, 786–798.
- Feller, D. *J. Comput. Chem.* **1996**, *17*, 1571–1586.
- Schuchardt, K. L.; Didier, B. T.; Elsethagen, T.; Sun, L.; Gurumoothi, V.; Chase, J.; Li, J.; Windus, T. L. *J. Chem. Inf. Model.* **2007**, *47*, 1045–1052.
- Ditchfield, R.; Hehre, W. J.; Pople, J. A. *J. Chem. Phys.* **1971**, *54*, 724–728.
- Mobley, D. L.; Chodera, J. D.; Dill, K. A. *J. Chem. Phys.* **2006**, *125*, 084902.
- Berendsen, H. J. C.; Postma, J. P. M.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

- (54) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.
- (55) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8592.
- (56) van der Spoel, D.; van Maaren, P. J. *J. Chem. Theory Comput.* **2006**, *2*, 1–11.
- (57) Lange, O. F.; van der Spoel, D.; de Groot, B. L. *Biophys. J.* **2010**, *99*, 647–655.
- (58) Nosé, S. *Mol. Phys.* **1984**, *52*, 255–268.
- (59) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695–1697.
- (60) Holian, B.; Voter, A.; Ravelo, R. *Phys. Rev. E* **1995**, *52*, 2338–2347.
- (61) Parrinello, M.; Rahman, A. *J. Appl. Phys.* **1981**, *52*, 7182–7190.
- (62) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J. Comput. Chem.* **1997**, *18*, 1463–1472.
- (63) Hess, B. *J. Chem. Theory Comput.* **2008**, *4*, 116–122.
- (64) van Gunsteren, W. F.; Berendsen, H. J. C. *Mol. Sim.* **1988**, *1*, 173–185.
- (65) van Gunsteren, W. F.; Berendsen, H. J. C. *Angew. Chem., Int. Ed. Engl.* **1990**, *29*, 992–1023.
- (66) Neumann, M. *Mol. Phys.* **1983**, *50*, 841–858.
- (67) Neumann, M.; Steinhäuser, O.; Pawley, G. S. *Mol. Phys.* **1984**, *52*, 97–113.
- (68) Lin, S. T.; Blanco, M.; Goddard, W. A., III. *J. Chem. Phys.* **2003**, *119*, 11792–11805.
- (69) Lin, S.-T.; Maiti, P. K.; Goddard, W. A., III. *J. Phys. Chem. B* **2010**, *114*, 8191–8198.
- (70) Pascal, T. A.; Lin, S.-T.; Goddard, W. A., III. *J. Phys. Chem. Chem. Phys.* **2011**, *13*, 169–181.
- (71) Berens, P. H.; Mackay, D. H. J.; White, G. M.; Wilson, K. R. *J. Chem. Phys.* **1983**, *79*, 2375–2389.
- (72) Hess, B. *J. Chem. Phys.* **2002**, *116*, 209–217.
- (73) Lide, D. R. *CRC Handbook of Chemistry and Physics*, 90th ed.; CRC Press: Cleveland, OH, 2009.
- (74) Frenkel, M.; Hong, X.; Dong, Q.; Yan, X.; Chirico, R. D. *Landolt-Börnstein - Group IV Physical Chemistry, Densities of Halohydrocarbons*; Springer: Berlin, 2000.
- (75) Yaws, C. L. *Thermophysical Properties of Chemicals and Hydrocarbons*; William Andrew Inc.: Beaumont, TX, 2008.
- (76) Caleman, C.; Hub, J. S.; van Maaren, P. J.; van der Spoel, D. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 6838–6842.
- (77) Wohlfahrt, C. *Landolt-Börnstein - Group IV Physical Chemistry*; Springer Verlag: New York, 2008; Chapter Static Dielectric Constants of Pure Liquids and Binary Liquid Mixtures. <http://www.springermaterials.com> (accessed Dec. 2011).
- (78) Whitfield, T. W.; Allison, G. J. M. S.; Bates, S. P.; Vass, H.; Crain, J. J. *J. Phys. Chem. B* **2005**, *110*, 3264–3673.
- (79) Jedlovsky, P.; Turi, L. *J. Phys. Chem. A* **1997**, *101*, 2662–2665.
- (80) MacCallum, J. L.; Tieleman, D. P. *J. Am. Chem. Soc.* **2002**, *124*, 15085–15093.
- (81) van der Spoel, D.; van Buuren, A. R.; Tieleman, D. P.; Berendsen, H. J. C. *J. Biomol. NMR* **1996**, *8*, 229–238.
- (82) Beachy, M.; Chasman, D.; Murphy, R.; Halgren, T.; Friesner, R. *J. Am. Chem. Soc.* **1997**, *119*, 5908–5920.
- (83) Lee, J.; Ripoll, D.; Czaplowski, C.; Pillardy, J.; Wedemeyer, W.; Scheraga, H. *J. Phys. Chem. B* **2001**, *105*, 7291–7298.
- (84) van der Spoel, D.; Lindahl, E. *J. Phys. Chem. B* **2003**, *107*, 11178–11187.
- (85) Lwin, T. Z.; Luo, R. *Protein Sci.* **2006**, *15*, 2642–2655.
- (86) Matthes, D.; de Groot, B. L. *Biophys. J.* **2009**, *97*, 599–608.
- (87) Jiang, J.; Wu, Y.; Wang, Z.-X.; Wu, C. *J. Chem. Theory Comput.* **2010**, *6*, 1199–1209.
- (88) Jha, S.; Coveney, P.; Laughton, C. *J. Comput. Chem.* **2005**, *26*, 1617–1627.
- (89) Jurecka, P.; Sponer, J.; Cerny, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985–1993.
- (90) Fadrna, E.; Spackova, N.; Sarzynska, J.; Koca, J.; Orozco, M.; Cheatham, T. E., III; Kulinski, T.; Sponer, J. *J. Chem. Theory Comput.* **2009**, *5*, 2514–2530.
- (91) Morgado, C. A.; Jurecka, P.; Svozil, D.; Hobza, P.; Sponer, J. *J. Chem. Theory Comput.* **2009**, *5*, 1524–1544.
- (92) Hemmingsen, L.; Madsen, D.; Esbensen, A.; Olsen, L.; Engelsen, S. *Carbohydr. Res.* **2004**, *339*, 937–948.
- (93) Hagler, A.; Lifson, S.; Dauber, P. *J. Am. Chem. Soc.* **1979**, *101*, 5122–5130.
- (94) Shirts, M. R.; Pande, V. S. *J. Chem. Phys.* **2005**, *122*, 134508.
- (95) White, B. R.; Wagner, C. R.; Truhlar, D. G.; Amin, E. A. *J. Chem. Theory Comput.* **2008**, *4*, 1718–1732.
- (96) Valdes, H.; Pluhackova, K.; Pitonak, M.; Rezac, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2747–2757.
- (97) Paton, R. S.; Goodman, J. M. *J. Chem. Inf. Model.* **2009**, *49*, 944–955.
- (98) Åqvist, J. *J. Phys. Chem.* **1990**, *94*, 8021–8024.
- (99) Hess, B.; Holm, C.; van der Vegt, N. *J. Chem. Phys.* **2006**, *124*, 164509.
- (100) Warren, G. L.; Patel, S. *J. Chem. Phys.* **2007**, *127*, 064509.
- (101) Mobley, D. L.; Barber, A. E., II; Fennell, C. J.; Dill, K. A. *J. Phys. Chem. B* **2008**, *112*, 2405–2414.
- (102) Fujitani, H.; Tanida, Y.; Ito, M.; Jayachandran, G.; Snow, C.; Shirts, M.; Sorin, E.; Pande, V. *J. Chem. Phys.* **2005**, *123*, 084108.
- (103) Felts, A.; Gallicchio, E.; Wallqvist, A.; Levy, R. *Proteins: Struct., Funct., Genet.* **2002**, *48*, 404–422.
- (104) Hetényi, C.; van der Spoel, D. *Protein Sci.* **2002**, *11*, 1729–1737.
- (105) Hetényi, C.; van der Spoel, D. *FEBS Lett.* **2006**, *580*, 1447–1450.
- (106) Hetényi, C.; van der Spoel, D. *Protein Sci.* **2011**, *20*, 880–893.
- (107) Industrial Fluid Properties Challenge. <http://www.fluidproperties.org> (accessed Dec. 2011).
- (108) Case, F. H.; Brennan, J.; Chaka, A.; Dobbs, K. D.; Friend, D. G.; Gordon, P. A.; Moore, J. D.; Mountain, R. D.; Olson, J. D.; Ross, R. B.; Schiller, M.; Shen, V. K.; Stahlberg, E. A. *Fluid Phase Equilib.* **2008**, *274*, 2–9.
- (109) McQuarrie, D. A. *Statistical Mechanics*; Harper & Row: New York, 1976.
- (110) Halgren, T. *J. Comput. Chem.* **1996**, *17*, 553–586.
- (111) Nevins, N.; Allinger, N. *J. Comput. Chem.* **1996**, *17*, 730–746.
- (112) Nevins, N.; Chen, K.; Allinger, N. *J. Comput. Chem.* **1996**, *17*, 669–694.
- (113) Nevins, N.; Lü, J.; Allinger, N. *J. Comput. Chem.* **1996**, *17*, 695–729.
- (114) Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A., Jr.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T. *J. Phys. Chem. B* **2010**, *114*, 2549–2564.
- (115) Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471–2474.
- (116) Kuo, L.; Mundy, C.; McGrath, M.; Siepmann, J.; VandeVondele, J.; Sprik, M.; Hutter, J.; Chen, B.; Klein, M.; Mohamed, F.; Krack, M.; Parrinello, M. *J. Phys. Chem. B* **2004**, *108*, 12990–12998.
- (117) Gageot, M. P. *J. Phys. Chem. A* **2008**, *112*, 13507–13517.
- (118) Cimas, A.; Vaden, T. D.; de Boer, T. S. J. A.; Snoek, L. C.; Gageot, M. P. *J. Chem. Theory Comput.* **2009**, *5*, 1068–1078.
- (119) Gageot, M.-P. *Phys. Chem. Chem. Phys.* **2010**, *12*, 3336–3359.
- (120) van der Spoel, D.; Hess, B. *Comput. Mol. Sci.* **2011**, *1*, 710–715.
- (121) Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Dill, K. A. *J. Phys. Chem. B* **2009**, *113*, 4533–4537.
- (122) Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Shirts, M. R.; Dill, K. A. *J. Chem. Theory Comput.* **2009**, *5*, 350–358.
- (123) Nicholls, A.; Mobley, D. L.; Guthrie, J. P.; Chodera, J. D.; Bayly, C. I.; Cooper, M. D.; Pande, V. S. *J. Med. Chem.* **2008**, *51*, 769–779.
- (124) Geballe, M. T.; Skillman, A. G.; Nicholls, A.; Guthrie, J. P.; Taylor, P. J. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 259–279.
- (125) van Maaren, P. J.; van der Spoel, D. *J. Phys. Chem. B* **2001**, *105*, 2618–2626.
- (126) Hill, E. A. *J. Am. Chem. Soc.* **1900**, *8*, 478–494.

Large-Scale MP2 Calculations on the Blue Gene Architecture Using the Fragment Molecular Orbital Method

Graham D. Fletcher,[†] Dmitri G. Fedorov,[‡] Spencer R. Pruitt,[§] Theresa L. Windus,[§] and Mark S. Gordon^{*,§}

[†]Argonne Leadership Computing Facility, Argonne, Illinois 60439, United States

[‡]National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan

[§]Iowa State University and Ames Laboratory, Ames, Iowa 50011, United States

ABSTRACT: Benchmark timings are presented for the fragment molecular orbital method on a Blue Gene/P computer. Algorithmic modifications that lead to enhanced performance on the Blue Gene/P architecture include strategies for the storage of fragment density matrices by process subgroups in the global address space. The computation of the atomic forces for a system with more than 3000 atoms and 44 000 basis functions, using second order perturbation theory and an augmented and polarized double- ζ basis set, takes ~ 7 min on 131 072 cores.

INTRODUCTION

In recent years, a breakdown in Dennard's scaling¹ has prevented CPU clock speeds from increasing significantly without giving rise to punitive power and cooling requirements. Fortunately, however, Moore's law² has continued to apply, and this has allowed processor designers to partially mitigate for the effect of plateauing clock speeds by having multiple compute units on a chip. This is evident in the mobile processor market where low power dual core processors are common all the way through to state of the art supercomputers equipped with massively multi-core processor nodes and special purpose accelerators. On the other hand, the aggregate floating point (or more general computational) performance of a chip/socket or even collection of chips/sockets in a node is expanding at a rate that is considerably greater than the speed at which data can be transferred between chips/sockets/nodes. This scenario significantly challenges the scalability of dense algebra problems, particularly the large, distributed matrix operations that are endemic in electronic structure algorithms.^{3,4} Scientific application programmers have in turn responded by seeking ways to break down a problem into manageable parts, exploiting the locality of certain properties to yield multiple levels of model scaling in a given method, as well as multiple levels of parallelism, in order to better map the computation onto the architecture. One such method in electronic structure theory is the fragment molecular orbital (FMO)^{5–7} method that is the focus of the work presented here.

While there are many other fragment-based methods,^{8–19} a distinctive feature of the FMO method is that the electrostatic potential (ESP) that represents the entire system is included during the calculation of the energy of each individual fragment. Further, a many-body expansion is used to account for the interfragment interactions. The FMO approach has the chief advantages of scaling nearly linearly in computation cost with the problem size, the avoidance of any empirically fitted parameters, and compatibility with all quantum chemical methods. Thus, the FMO method offers considerable flexibility to mitigate the traditional bottlenecks of quantum chemistry in terms of cost, memory, and communication bandwidth. Parallel scalability derives from the

concurrent and asynchronous execution of individual fragment calculations on distinct processor subgroups, provided that even load balancing can be achieved as it has been in this work.

For parallel execution, the FMO implementation in GAMESS (General Atomic and Molecular Electronic Structure System)^{20,21} can make use of the distributed data interface (DDI),^{22,23} and its generalization to subgroups, the generalized DDI (GDDI).²⁴ GDDI allows processor subgroups to be created in such a way that, during an FMO calculation, they can access fragment data from other subgroups asynchronously in a "one-sided" fashion. Consequently, the FMO method is emerging as a highly effective means of harnessing modern supercomputing hardware to treat systems with thousands of atoms quantum-mechanically.

The FMO method has been applied to a broad range of large systems. Among many important examples of FMO applications are studies in protein folding²⁵ and drug design,^{26–29} and an interface with the qualitative structure–activity relationship (QSAR).³⁰ The FMO method has also been applied to oligosaccharides,³¹ zeolites,³² nanowires,³³ and molecular clusters, in particular to the explicit treatment of solvents.^{34,35}

One purpose of the present study is to demonstrate that the FMO method can make effective use of massively parallel computers that approach the petascale (i.e., the effective use of $\sim 100\,000$ or more compute cores) in both speed and resources. Second, this capability will facilitate future applications for the study of large, complex chemical problems that might otherwise be computationally intractable. Software development for massively parallel (i.e., peta and exascale) computers is frequently well behind the advances in hardware; therefore, efficient computational methods are needed to take advantage of the new computational architectures that are becoming available, as well as those that are anticipated in the near future. The FMO method discussed in the present work is one viable example of new high performance software in electronic structure theory.

Received: August 7, 2011

Published: December 12, 2011

METHODS AND COMPUTATIONAL DETAILS

The FMO method has been described extensively elsewhere;^{6,7,36} therefore, only a brief description will be given here. The basic equation to obtain the total energy of the system divided into N fragments is

$$E^{\text{FMO2}} = \sum_I E'_I + \sum_{I>J} (E'_{IJ} - E'_I - E'_J) + \sum_{I>J} \text{Tr}(\Delta\mathbf{D}^{IJ}\mathbf{V}^{IJ}) \quad (1)$$

where E'_I and E'_{IJ} are the internal energies of fragments I (monomers) and their pairs IJ (dimers), polarized by the ESP field of all other fragments determined self-consistently. $\Delta\mathbf{D}^{IJ}$ is the difference between the density of the dimer IJ and the sum of the densities of the monomers I and J ; \mathbf{V}^{IJ} is the ESP for dimer IJ . The FMO2 level of theory includes the explicit pair corrections shown in the second and third terms of eq 1, while FMO1 corresponds to the sum over monomers in the first term of the equation. The gradient is obtained by taking the fully analytic derivative of the FMO energy, recently developed by Nagata et al.^{37,38} The FMO1 method, which was used for scalability tests in the previous papers, gives the internal energies of the fragments in the presence of the fields of all other fragments and also describes the many-body polarization, as demonstrated by the pair interaction energy decomposition analysis (PIEDA).³⁹ The individual fragment polarization energies E_I^{PL} can be calculated as

$$\Delta E_I^{\text{PL}} = - (E'_I - E_I^0) \quad (2)$$

where E_I^0 is the internal energy of fragment I (i.e., the energy computed without the field of the other fragments). All of the energies E_I^0 can be calculated in a single run and used to estimate the many-body polarization in large systems.

The key to achieving high parallel efficiency is the multilevel hierarchical approach, GDDI, that has so far been limited to two levels.^{24,40} Specifically, computer nodes are divided into groups, and each group is assigned a particular fragment or fragment pair calculation to perform. The calculations within a group are performed without communication to other groups. However, at several points during the calculation some communications are required. The most important of these communications is to exchange fragment densities and to accumulate a total property, such as the energy. All workload balancing is done dynamically, at both the inter- and intragroup levels.

In general, the number of groups is chosen to balance the losses due to synchronization. For example, when a group finishes early and has to wait for the others to finish, having fewer groups is more efficient. On the other hand, when fewer nodes are assigned to a group, the parallelization within a group (e.g., an RHF or MP2 calculation of a given fragment or dimer) is more efficient, so having more groups is preferred. It has been shown previously²⁴ how varying the group size affects the synchronization and data exchange timings, as well as the general performance. For water clusters with a uniform fragment size, the load balancing is simpler than for proteins,²⁴ for which other techniques such as doing the larger jobs first²⁴ and employing semidynamic load balancing⁴⁰ (i.e., static load balancing on large GDDI groups for a few large fragments and dynamic load balancing on small GDDI groups for the rest) have been found to be necessary. Because the number of tasks is different for monomers vs dimers, a different grouping strategy is often employed for these two different components of a calculation, with the regrouping performed at the end of the monomer step. Guidelines for the grouping of CPU cores have been given elsewhere.⁹ In general, it is recommended that each

group does several calculations for more even load balancing, especially when fragments are different sizes. As an example, consider FMO2 calculations on 4096 water molecules at the RHF/6-31G(d) level of theory, with each fragment defined to be two water molecules. This means there are 2 098 128 dimers. Most of these dimers are far enough apart to be treated using the ESP approximation, but 24 411 dimers must still be treated with quantum mechanics. If one uses 2048 groups for the dimer calculations, there will be ~ 12 dimers/group. The total wall clock time for this set of calculations is 11.5 min. Doubling the number of dimer groups to 4096 increases the wall clock time to 14.6 min. So, there is clearly a dependence on the number of groups that is used in each step of a calculation. In the present work, the default FMO options are used, except that all ESPs are computed using the one-electron approximation.

Prior to this work, FMO calculations were performed on computer clusters with dozens, or at most hundreds, of CPU cores with local disks attached to each node. In such cases, the I/O required to store and access the fragment densities is negligible compared to the total times. However, on supercomputers containing thousands to tens of thousands of cores, the I/O constitutes a major bottleneck. Due to recent multicore CPU development, the CPU core count continues to increase while the number and efficiency of storage devices typically lags far behind. Additionally, many modern massively parallel computers such as the K computer⁴¹ or the Blue Gene/P computer⁴² have no local storage. The I/O system that is usually provided on such computers generally does not meet the I/O demands in the FMO production code, due to the required access to fragment densities in order to calculate ESPs.

In previous FMO implementations, the master process of each group created a direct-access file in which the densities of all fragments are stored. The new approach, employed here, is based on a large array containing all fragment densities created in shared memory distributed among nodes. The standard DDI functionality described elsewhere is used.²³ The fragment densities are stored on data servers and sent on demand to compute nodes directly, with these communications sometimes involving intergroup operations. In addition, the RAMDISK feature on the Blue Gene/P was found to be effective in reducing I/O to hard disks by causing scratch files to be written to the memory of the nearest I/O node, thereby greatly improving performance. The main difference between the density file and other scratch files is that the latter are local to each group calculation and need not be made accessible to other groups, while the former must be made either fully global or synchronized between groups at some points. The underlying MP2 gradient calculations also use DDI memory.^{22,43}

As an example of the consequent gain in efficiency, consider 1024 water molecules whose energy was calculated using FMO2 with MP2 and the 6-31G(d,p) basis set, both with the previous disk-based implementation ("FMOd") and the new implementation ("FMOm"). Each calculation was run on 1024 nodes (4096 cores) on a Blue Gene/P computer. The wall time required for the FMOd calculation was 335.4 min, whereas the corresponding FMOm wall time was 10.7 min. This is a dramatic improvement in efficiency. This 31-fold speed-up demonstrates that the DDI-based density storage is paramount to running FMO calculations effectively on large-scale parallel computers.

The hardware platform used for all calculations was the Blue Gene/P computer (*Intrepid*) at the Argonne Leadership Computing Facility (ALCF). A Blue Gene/P node consists of a quad-core

Table 1. The Performance of FMO2-MP2 Force Calculations on the Blue Gene/P^a

		racks:	1	2	4	8	16	32
		cores:	4096	8192	16 384	32 768	65 536	131 072
waters	atoms	basis functions	wall time (min)					
128	384	2432	0.5	0.4				
256	768	4864	1.1	0.7	0.5			
512	1536	9728	3.6	2.0	1.2	0.8	0.7	
1024	3072	19 456	10.7	6.3	3.4	2.1	1.5	1.2
2048	6144	38 912	34.6	18.5	11.1	6.1	3.7	2.6

^aThe atomic basis set is 6-31G(d).

Table 2. The Performance of FMO2-MP2 Force Calculations on the Blue Gene/P^a

		racks:	1	2	4	8	16	32
		cores:	4096	8192	16 384	32 768	65 536	131 072
waters	atoms	basis functions	wall time (min)					
128	384	5504	8.6	4.8	2.7	1.8		
256	768	11 008	19.8	10.5	5.8	3.4	2.2	
512	1536	22 016		28.9	15.4	8.6	4.9	3.2
1024	3072	44 032			41.1	22.0	12.2	7.1

^aThe atomic basis set is aug-cc-pVDZ.

PPC450 chip, running at 850 MHz, with 2 GB of DRAM. Each board contains 32 compute nodes and has two dedicated I/O nodes to handle access to disk. The relatively low clock rate and node memory combine to give the architecture a very desirable FLOP/Watt ratio of 360 MFLOPS/Watt. Nodes are linked via three networks, one for collective communications with a bandwidth of 6.8 GB/s, a point-to-point interconnect of 3D torus topology with a bandwidth of 3.4 GB/s, and a 10 GB/s Ethernet link for I/O. Intrepid has 40 “racks” of 1024 nodes each, giving it a total of 163 840 processor cores.

RESULTS

Water clusters were used as the test systems for the benchmarks that are presented here, as they permit a convenient and systematic series of problem sizes. The following procedure was used to construct the water clusters used in this work. First, the oxygen atoms were fixed to grid points with an increment equivalent to the O–O separation of a typical hydrogen bond, taken to be 2.98 Å. The hydrogen atoms were fixed at the MP2/aug-cc-pVDZ O–H bond length and angle, matching the intended level of theory that will be subsequently applied to the entire cluster. The directional orientation of this fixed geometry was chosen at random. Water molecules were first arranged in cubes containing eight waters. This arrangement yields fragments containing 1, 2, 4, and 8 waters in a “droplet” configuration rather than a “chain”, since the former is more favorable to the convergence of both the local and global FMO-RHF equations, and since bulk water contains primarily droplet-like clusters. A sufficient number of cubes is then created in order to yield the desired “slab” of water, itself as close to cubic as possible.

To survey a range of cluster sizes, calculations were performed on clusters containing 128, 256, 512, 1024, 2048, and 4096 waters. To assess scalability across a range of processor partitions, 1, 2, 4, 8, 16, and 32 racks of Intrepid were used. In all calculations, the FMO series is taken to second-order (FMO2),

sufficient for kcal/mol agreement of the energies with full calculations. The choice of two water molecules per fragment was found to give kcal/mol accuracy as well as the best overall performance, scalability, and execution time. The number of processor groups was chosen to be equal to the number of fragments. Larger choices for the number of groups, as might be done in an attempt to align that number with the number of so-called “self-consistent” dimers, were found to adversely increase the communication overhead and give a sharp decline in performance. The level of theory was chosen to be MP2, because (despite its limitations) MP2 has been found to give results in good agreement with coupled cluster, CCSD(T), calculations for water clusters.⁴⁴

Single-point energies and gradients were computed for all clusters. The gradients will be used in future dynamics simulations. Two series of calculations were performed spanning the cluster sizes and processor partitions noted above: the first series uses the 6-31G(d) atomic basis set⁴⁵ (see Table 1); the second series uses the aug-cc-pVDZ basis set⁴⁶ (see Table 2). The next basis set in the “Dunning” series, aug-cc-pVTZ,⁴⁶ is more than twice as large (105 Cartesian basis functions for water) as the aug-cc-pVDZ basis set (43 Cartesian functions), so only a short-range of small clusters could be computed in a reasonable time with the current hardware. Calculations in a given partition series (row) of the tables begin below 1 h in duration and continue to either the point of “diminishing returns” regarding scalability or the maximum partition, whichever comes first. The scalability of the largest calculation, that of 4096 waters (having 12 288 atoms) at the MP2/6-31G(d) level of theory, is depicted in Figure 1, where it may be seen that the computations scale rather well with a system size up to the full complement of more than 131 000 cores.

The scaling of the cost of the calculation with problem size may be assessed by examining the columns in Tables 1 and 2. In Table 1, with the smaller basis set, doubling the size of the

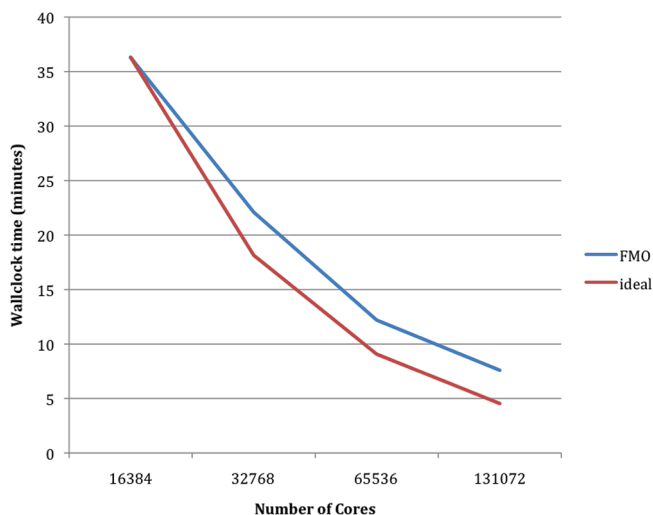


Figure 1. FMO2-MP2/6-31G(d) forces calculation of 12 288 atoms on Blue Gene/P.

system increases the cost by roughly a factor of 3. For the larger basis set, the corresponding factor is a bit smaller, between 2 and 3. This is not as good as linear scaling, but considerably better than quadratic scaling. The deviation from the linear regime is mainly caused by the increasing fraction of the far separated dimers computed with the electrostatic approximation. As the cluster size increases, the parallel scalability with the number of cores improves, as one would expect. The aug-cc-pVDZ calculations are on the order of 10 times more expensive than the corresponding 6-31G(d) calculations. This is consistent with the basis set roughly doubling in size on going from the smaller (Table 1) to the larger (Table 2) basis set, combined with the cubic increase in the fragment cost with the number of basis functions. Of great interest with regard to the use of the FMO method are the prospects for *ab initio* MP2 molecular dynamics (MD) simulations of bulk water and other fluids. The approximate linear scaling exhibited in the tables makes such calculations feasible. The present benchmarks provide reference points for the estimation of costs within the current hardware constraints. For example, MD step times on the order of 1 min for clusters 512–1024 water molecules are highly encouraging.

SUMMARY

This work has demonstrated that the FMO code in GAMESS can efficiently utilize “petascale” processor counts. Specifically, the GAMESS/FMO method can use up to 131 072 processor cores of a Blue Gene/P supercomputer to obtain the MP2 atomic forces for a system with more than 4000 atoms, with the 6-31G(d) basis set, in approximately 7 min. Of course, the Blue Gene/P is not the only supercomputer architecture capable of achieving petascale computing. However, the group division of processors chosen for this work is unique to the Blue Gene/P architecture due to the great number of variables required for optimal efficiency during electronic structure calculations. These variables include fragment size (number of basis functions per fragment and homogeneity of fragment size), level of theory (e.g., HF, MP2 etc), and most importantly the computer architecture being used. While the first two variables can affect the accuracy of the FMO calculation, the last variable can also affect these choices due to computational resources available.

Likewise, depending on the level of theory, memory requirements may necessitate larger GDDI groups.

The number of cores per node, amount of memory per core, network speed, and CPU clock rate as well as the choice of communication layer (e.g, MPI, ARMC) can all affect how the user chooses the GDDI group division. The relatively small node size (four cores per node) of the Blue Gene/P architecture makes it particularly flexible in terms of the GDDI group size. However, other architectures, such as the Cray XE6 and SGI Altix, typically consist of 12–16 cores per node with significantly higher clock rates. These numbers are only likely to increase with advances in microprocessor technology, leading to an increase in the minimum GDDI group size. The obvious solution to this issue is to simply increase the fragment size in order to maintain a high level of computational efficiency, with the byproduct being an increase in the accuracy of the FMO calculation. This exemplifies the importance of GDDI group choice by showing the importance of proper group size for the appropriate fragment size.

Problem sizes in the thousands-of-atoms range represent a “critical mass” in the applicability of electronic structure theory to chemistry at which many issues of national strategic importance (for instance, renewable energy and medicine), including systems on the biological scale (e.g., proteins), become accessible with predictive capability and detailed understanding. The FMO method is clearly able to address such problems with both efficiency and accuracy. Further improvements in performance will necessitate addressing several remaining bottlenecks. Among these are load balancing, the communications (e.g., input/output) overhead, and the use of outdated programming paradigms, especially in legacy codes. All of these issues are being addressed by the authors and colleagues and will continue to be addressed in the future.

AUTHOR INFORMATION

Corresponding Author

*E-mail: mark@si.msg.chem.iastate.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

This research used resources of the Argonne Leadership Computing Facility at Argonne National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under contract DE-AC02-06CH11357, made possible by a Department of Energy INCITE grant. The research was supported by a Department of Energy PCTC grant, and by the Air Force Office of Scientific Research, both to M.S.G. Funding for T.L.W. has been provided by an NSF grant for petascale applications. D.G.F. thanks Drs. H. Shitara, H. Umeda, and Y. Itono for fruitful discussions about FMO parallelization and the Next Generation SuperComputing Project, Nanoscience Program (MEXT, Japan) for partial financial support. The authors are most grateful to Professor Alistair Rendell for his insightful comments.

REFERENCES

- (1) The Impact of Dennard's Scaling Theory. *IEEE Solid-State Circuits Society News*, **2007**, *12* (1).

- (2) Moore, G. E. Cramming More Components Onto Integrated Circuits. *Electronics* **1965**, *38*, 114–117.
- (3) Ishimura, K.; Kuramoto, K.; Ikuta, Y.; Hyodo, S. MPI/OpenMP Hybrid Parallel Algorithm for Hartree–Fock Calculations. *J. Chem. Theory Comput.* **2010**, *6*, 1075–1080.
- (4) Aprà, E.; Harrison, R. J.; Shelton, W. A.; Tipparaju, V.; Vázquez-Mayagoitia, A. Computational chemistry at the petascale: are we there yet? *J. Phys. Conf. Ser.* **2009**, *180*, 012027.
- (5) Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M. Fragment molecular orbital method: an approximate computational method for large molecules. *Chem. Phys. Lett.* **1999**, *313*, 701–706.
- (6) Fedorov, D. G.; Kitaura, K. Extending the Power of Quantum Chemistry to Large Systems with the Fragment Molecular Orbital Method. *J. Phys. Chem. A* **2007**, *111*, 6904–6914.
- (7) Fedorov, D. G.; Kitaura, K. *The Fragment Molecular Orbital Method: Practical Applications to Large Molecular Systems*; CRC Press: Boca Raton, FL, 2009.
- (8) Otto, P. Investigation of the stability of oligonucleotides and oligodinucleotides. *THEOCHEM* **1989**, *188*, 277–288.
- (9) Gao, J. L. Toward a molecular orbital derived empirical potential for liquid simulations. *J. Phys. Chem. B* **1997**, *101*, 657–663.
- (10) Sorkin, A.; Dahlke, E. E.; Truhlar, D. G. Application of the Electrostatically Embedded Many-Body Expansion to Microsolvation of Ammonia in Water Clusters. *J. Chem. Theory Comput.* **2008**, *4*, 683–688.
- (11) Söderhjelm, P.; Ryde, U. How Accurate Can a Force Field Become? A Polarizable Multipole Model Combined with Fragment-wise Quantum-Mechanical Calculations. *J. Phys. Chem. A* **2009**, *113*, 617–627.
- (12) Gordon, M. S.; Mullin, J. M.; Pruitt, S. R.; Roskop, L. B.; Slipchenko, L. V.; Boatz, J. A. Accurate Methods for Large Molecular Systems. *J. Phys. Chem. B* **2009**, *113*, 9646–9663.
- (13) Mata, R. A.; Stoll, H.; Cabral, B. J. C. A Simple One-Body Approach to the Calculation of the First Electronic Absorption Band of Water. *J. Chem. Theory Comput.* **2009**, *5*, 1829–1837.
- (14) Weiss, S. N.; Huang, L.; Massa, L. A generalized higher order kernel energy approximation method. *J. Comput. Chem.* **2010**, *31*, 2889–2899.
- (15) Hua, S. G.; Hua, W. J.; Li, S. H. An Efficient Implementation of the Generalized Energy-Based Fragmentation Approach for General Large Molecules. *J. Phys. Chem. A* **2010**, *114*, 8126–8134.
- (16) Řežáč, J.; Salahub, D. R. Multilevel Fragment-Based Approach (MFBA): A Novel Hybrid Computational Method for the Study of Large Molecules. *J. Chem. Theory Comput.* **2010**, *6*, 91–99.
- (17) Yeole, S. D.; Gadre, S. R. On the applicability of fragmentation methods to conjugated π systems within density functional framework. *J. Chem. Phys.* **2010**, *132*, 094102.
- (18) He, X.; Merz, K. M., Jr. Divide and Conquer Hartree–Fock Calculations on Proteins. *J. Chem. Theory Comput.* **2010**, *6*, 405–411.
- (19) Kobayashi, M.; Kunisada, T.; Akama, T.; Sakura, D.; Nakai, H. Reconsidering an analytical gradient expression within a divide-and-conquer self-consistent field approach: Exact formula and its approximate treatment. *J. Chem. Phys.* **2011**, *134*, 034105.
- (20) Schmidt, M. W.; Baldrige, 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. General atomic and molecular electronic structure system. *J. Comput. Chem.* **1993**, *14*, 1347–1363. <http://www.msg.ameslab.gov/gamess/index.html> (accessed Dec. 2011).
- (21) Gordon, M. S.; Schmidt, M. W. Advances in electronic structure theory: GAMESS a decade later. In *Theory and Applications of Computational Chemistry, the first forty years*; Elsevier: Amsterdam, 2005; pp 1167–1189.
- (22) Fletcher, G. D.; Gordon, M. S.; Schmidt, M. W. Developments in Parallel Electronic Structure Theory. *Adv. Chem. Phys.* **1999**, *110*, 267.
- (23) Fletcher, G. D.; Schmidt, M. W.; Bode, B. M.; Gordon, M. S. The Distributed Data Interface in GAMESS. *Comput. Phys. Commun.* **2000**, *128*, 190–200.
- (24) Fedorov, D. G.; Olson, R. M.; Kitaura, K.; Gordon, M. S.; Koseki, S. A new hierarchical parallelization scheme: generalized distributed data interface (GDDI), and an application to the fragment molecular orbital method (FMO). *J. Comput. Chem.* **2004**, *25*, 872–880.
- (25) He, X.; Fusti-Molnar, L.; Cui, G.; Merz, K. M., Jr. Importance of dispersion and electron correlation in ab initio protein folding. *J. Phys. Chem. B* **2009**, *113*, 5290–5300.
- (26) Sawada, T.; Fedorov, D. G.; Kitaura, K. Role of the key mutation in the selective binding of avian and human influenza hemagglutinin to sialosides revealed by quantum-mechanical calculations. *J. Am. Chem. Soc.* **2010**, *132*, 16862–16872.
- (27) Fujimura, K.; Sasabuchi, Y. The Role of Fluorine Atoms in a Fluorinated Prostaglandin Agonist. *Chem. Med. Chem.* **2010**, *5*, 1254–1257.
- (28) Ohno, K.; Mori, K.; Orita, M.; Takeuchi, M. Computational Insights into Binding of Bisphosphates to Farnesyl Pyrophosphate Synthase. *Curr. Med. Chem.* **2011**, *18*, 220–233.
- (29) Fedorov, D. G.; Alexeev, Y.; Kitaura, K. Geometry Optimization of the Active Site of a Large System with the Fragment Molecular Orbital Method. *J. Phys. Chem. Lett.* **2011**, *2*, 282–288.
- (30) Mazanetz, M. P.; Ichihara, O.; Law, R. J.; Whittaker, M. Prediction of cyclin-dependent kinase 2 inhibitor potency using the fragment molecular orbital method. *J. Cheminformatics* **2011**, *3*, 2.
- (31) Sawada, T.; Fedorov, D. G.; Kitaura, K. Structural and interaction analysis of helical heparin oligosaccharides with the fragment molecular orbital method. *Int. J. Quantum Chem.* **2009**, *109*, 2033–2045.
- (32) Fedorov, D. G.; Jensen, J. H.; Deka, R. C.; Kitaura, K. Covalent Bond Fragmentation Suitable To Describe Solids in the Fragment Molecular Orbital Method. *J. Phys. Chem. A* **2008**, *112*, 11808–11816.
- (33) Fedorov, D. G.; Avramov, P. V.; Jensen, J. H.; Kitaura, K. Analytic gradient for the adaptive frozen orbital bond detachment in the fragment molecular orbital method. *Chem. Phys. Lett.* **2009**, *477*, 169–175.
- (34) Sato, M.; Yamataka, H.; Komeiji, Y.; Mochizuki, Y.; Ishikawa, T.; Nakano, T. How Does an S_N^2 Reaction Take Place in Solution? Full Ab Initio MD Simulations for the Hydrolysis of the Methyl Diazonium Ion. *J. Am. Chem. Soc.* **2008**, *130*, 2396–2397.
- (35) Kistler, K. A.; Matsika, S. Solvatochromic Shifts of Uracil and Cytosine Using a Combined Multireference Configuration Interaction/Molecular Dynamics Approach and the Fragment Molecular Orbital Method. *J. Phys. Chem. A* **2009**, *113*, 12396–12403.
- (36) Nagata, T.; Fedorov, D. G.; Kitaura, K. In *Linear-Scaling Techniques in Computational Chemistry and Physics*; Zalesny, R., Papadopoulos, M. G., Mezey, P. G., Leszczynski, J., Eds.; Springer: Berlin, 2011; Chapter 2.
- (37) Nagata, T.; Fedorov, D. G.; Kitaura, K. Derivatives of the approximated electrostatic potentials in the fragment molecular orbital method. *Chem. Phys. Lett.* **2009**, *475*, 124–131.
- (38) Nagata, T.; Fedorov, D. G.; Kitaura, K. Importance of the hybrid orbital operator derivative term for the energy gradient in the fragment molecular orbital method. *Chem. Phys. Lett.* **2010**, *492*, 302–308.
- (39) Fedorov, D. G.; Kitaura, K. Pair interaction energy decomposition analysis. *J. Comput. Chem.* **2007**, *28*, 222–237.
- (40) Ikegami, T.; Ishida, T.; Fedorov, D. G.; Kitaura, K.; Inadomi, Y.; Umeda, H.; Yokokawa, M.; Sekiguchi, S. Full Electron Calculation Beyond 20,000 Atoms: Ground Electronic State of Photosynthetic Proteins. *Proc. of Supercomputing 2005*; IEEE Computer Society: Seattle, 2005.
- (41) See, for example, <http://www.fujitsu.com/global/news/pr/archives/month/2011/20110620-02.html>.
- (42) See, for example, http://en.wikipedia.org/wiki/Blue_Gene#Blue_Gene.2FP (accessed Dec. 2011).
- (43) Fletcher, G. D.; Rendell, A. P.; Sherwood, P. A Parallel Second-Order Møller-Plesset Gradient. *Mol. Phys.* **1997**, *91*, 431–438.
- (44) Xantheas, S. S.; Burnham, C. J.; Harrison, R. J. *J. Chem. Phys.* **2002**, *116*, 1493–1499. Olson, R. M.; Bentz, J. L.; Kendall, R. A.; Schmidt, M. W.; Gordon, M. S. *J. Comput. Theor. Chem.* **2007**, *3*, 1312–1328.
- (45) Hehre, W. J.; Ditchfield, R.; Pople, J. A. *J. Chem. Phys.* **1972**, *56*, 2257–2261.
- (46) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007–1023.

Performance of Effective Core Potentials for Density Functional Calculations on 3d Transition Metals

Xuefei Xu[†] and Donald G. Truhlar^{*,†}[†]Department of Chemistry and Supercomputing Institute, University of Minnesota, Minneapolis, Minnesota 55455-0431, United States

Supporting Information

ABSTRACT: The performance of popular Hartree–Fock-based effective core potentials in Hartree–Fock and density functional calculations of 3d transition metals has been evaluated by basis-set convergence studies for ten cases: the equilibrium bond dissociation energy (D_e) for dissociation of ground-state Ti_2 to ground and excited atoms, the ground-state dissociation energies of FeO , Cu_2 , ScH , TiH , Sc_2 , Fe_2 , and TiV^+ , and the first excitation energy (E_x) of Ti atom. Each case is studied with 11 or 13 density functionals. For comparison, the accuracy of the all-electron def2-TZVP basis set is tested with both relativistic and nonrelativistic treatments. Convergence and accuracy are assessed by comparing to relativistic all-electron calculations with a nearly complete relativistic basis set (NCBS-DK, which denotes the cc-pVSZ-DK basis set for 3d metals and hydrogen and the ma-cc-pVSZ-DK basis set for oxygen) and to nonrelativistic all-electron calculations with a nearly complete nonrelativistic basis set (NCBS-NR, which denotes the cc-pVSZ basis set for 3d metals and hydrogen and the ma-cc-pVSZ basis set for oxygen). As compared to NCBS-DK results, all ECP calculations perform worse than def2-TZVP all-electron relativistic calculations when averaged over all 130 data (13 functionals and ten test cases). The compact effective potential (CEP) relativistic effective core potential (RECP) combined with a valence basis set developed for the many-electron Dirac–Fock (MDF10) RECP performs best in effective core potential calculations and has an average basis-set incompleteness error of 3.7 kcal/mol, which is much larger than that (0.9 kcal/mol) of def2-TZVP relativistic all-electron results. Hence, the def2-TZVP relativistic all-electron calculations are recommended for accurate DFT calculations on 3d transition metals. In addition to our general findings, we observed that all kinds of density functionals do not show the same trends. For example, when ECPs are used with hybrid functionals, which sometimes are not recommended for calculations of transition metal systems, they are found to perform better at achieving the basis-set limit than when used with local functionals and meta-GGA functionals. The most successful combination of RECP and basis set has a basis-set incompleteness error of 1.7–2.4 kcal/mol for hybrid generalized gradient approximations, which is smaller than that of nonrelativistic NCBS calculations (whose average basis-set incompleteness error for hybrid functionals is 2.7–2.9 kcal/mol). The average basis-set incompleteness error in Hartree–Fock calculations is 1.0–4.4 kcal/mol for five of the ECP basis sets but is 5.8–10.8 kcal/mol for six others.

1. INTRODUCTION

Accurate computational study of transition metal (TM) containing systems has been recognized as a challenge for several decades because of the large computational cost, significant relativistic effects, and particularly the presence of low-lying electronic states arising from partially filled d shells. Because of the complexity, the present most widely used computational method for large TM system is density functional theory (DFT) combined with the use of effective core potentials (ECPs).¹ Following this strategy, there are several issues to which one should pay attention.

The first issue is the decision whether to use an ECP. An ECP is a potential energy function added to an electronic structure calculation to replace the explicit treatment of core electrons. The use of an ECP allows one to significantly reduce the cost of calculations in three ways: (i) by decreasing the number of basis functions (because no basis functions are required to treat core orbitals), (ii) by decreasing basis-set superposition error; and (iii) by solving a nonrelativistic (NR) wave equation for valence orbitals in the presence of a relativistic ECP (RECP, which is an ECP fitted to relativistic calculations) rather than having to use a relativistic all-electron (AE) wave equation to include the relativistic effects. The second issue is that ECP requirements for DFT are different from those for wave function theory (WFT).

The most popular ECPs for DFT calculations with Gaussian basis functions have been developed with Hartree–Fock (HF) wave functions, and their application in DFT studies needs further validation because of the nonlinear dependence of exchange–correlation (xc) functional on density. The importance of nonlinear core corrections (NLCC) in DFT calculations with ECPs has been stressed in several studies,² and a few systematic investigations³ have been carried out to validate the use of some HF-derived ECPs in Gaussian-based DFT calculations. In our recent work on ECPs for As-containing compounds, we investigated⁴ the need for NLCCs for Gaussian-based DFT calculations. The limited transferability of HF-derived ECPs in DFT calculations was observed, but the errors introduced relative to the nearly complete all-electron (AE) basis set calculations by using small-core ECPs with appropriate valence basis sets for arsenic were found to be small enough to be acceptable for many purposes. However, for transition metals, the conclusions could change, especially due to the variability of s and d orbital occupations upon bonding or electronic excitation. With different valence occupancies, the electron density in the

Received: August 11, 2011

Published: October 20, 2011

core region of a particular TM atom could differ significantly from the density of the reference state used in the ECP construction; this makes stronger demands on the transferability of ECPs for accurate calculations with either HF or DFT.²

A related issue is that ECPs do not eliminate electron density in the core region. They do remove the density due to core orbitals, and they replace the core-region density of the true core orbitals by the core-region density of the pseudovalence orbitals (which are the nodeless valence orbitals determined in the presence of ECPs rather than in the presence of explicit core orbitals). Different ECPs change the core-region density and density gradient in different ways, and different density functionals may be more or less suitable for use in the presence of such changes as caused by a particular ECP. Meta-GGA functionals that involve orbital-dependent kinetic energy density make the importance of this complication even more difficult to pre-empt.

Another issue is that, despite the success of DFT for systems composed of main-group elements, density functionals have special problems for the investigation of TM systems.⁵ The main one is that, as a single reference method, DFT is expected to be less accurate for some TM systems, because of their strong multireference character. The performance of various density functionals when applied to TM-containing systems has been extensively investigated using AE basis sets in the past decade.⁶ One conclusion^{6b} is that local generalized gradient approximations (GGA) and meta-GGA functionals usually perform better for TM species than do the nonlocal hybrid functionals that are the recommended functionals for most main-group species; this results because the Hartree–Fock (HF) exchange involved in hybrid functionals does not incorporate the important static correlation effects of multireference systems, whereas density functional exchange includes a portion of the nondynamical correlation energy,⁷ this is particularly important in TM bonding.⁸ The various exchange and correlation functionals include different approximations to Hartree–Fock exchange, nondynamical correlation, and dynamical correlation, and the resulting errors introduced by each density functional interact nonlinearly with errors introduced by each ECP, by the NLCC, and by the pseudovalence orbital density. Ultimately we need to find an ECP that works well for a given density functional.

In the present work, we investigate the performance of popular HF-derived ECPs for DFT study of 3d TM-containing species. A key issue in the construction of effective core potentials is that they do not represent just the interaction of a single electron with the core. Rather they must be constructed from the neutral atom wave function so that they take account, in an effective way, of the valence–valence interaction energy between the pseudovalence orbitals without modifying the form of the two-electron Coulomb interaction that is appropriate for interactions among the original valence orbitals.⁹ This indicates that effective core potential cannot be completely transferable among molecules that have different valence electron distributions or different partial atomic charges. Therefore, in testing the performance of effective core potentials we should consider molecules with bonds having both high and low partial ionic character, and we should consider molecules in which the valence electron distribution is significantly different than that in the atoms, for example a case where s and d orbital occupancies have changed. The test set used here includes this kind of diversity. The first and major part of the present investigation involves determining how well one can find a generally applicable ECP for DFT methods;

this is accomplished by averaging ECP performance on up to 130 data (as explained above). The second part is a study of whether particular ECPs are well suited for use with specific density functionals. The appropriate valence basis sets to be combined with ECPs for DFT calculations are also discussed. In addition we carry out HF calculations for sorting out the special ECP needs of DFT as compared to those of WFT.

2. COMPUTATIONAL DETAILS

It is well-known that core–valence interactions are too strong to allow one to replace the 3s and 3p electrons by an ECP in 3d TMs.¹⁰ Therefore, we consider only “small-core” ECPs that replace the innermost 10 electrons. Finally we note that the relativistic effects considered in the present article are only the one-electron scalar relativistic effects, that is, the mass-velocity and Darwin terms; we do not consider spin–orbit coupling at all.

Our investigation starts by calculating the equilibrium bond energy (D_e) of Ti_2 using a variety of combinations of popular ECPs and valence basis sets and 11 popular and prototypical functionals (M05,¹¹ M06-L,¹² M06,¹³ BLYP,¹⁴ ω B97X-D,¹⁵ τ HCTHhyb,¹⁶ G96LYP,^{14b,17} mPWLYP,^{14b,18} B3LYP,^{14,19} X3LYP,²⁰ and BPBE^{14a,21}). We investigate the $^3\Delta_g$ state of Ti_2 and consider the two dissociation limits of this state: one dissociation limit is the ground state (3F , $4s^23d^2$) of two Ti atoms; another dissociation limit is the first excited state (5F , $4s^13d^3$) of two Ti atoms; the former may be called adiabatic dissociation, and the latter may be called diabatic dissociation because the bonding of Ti_2 ($^3\Delta_g$) is derived from the $4s^13d^3$ occupations of Ti atoms.

For each kind of dissociation and each given functional, we carry out a Douglas–Kroll–Hess second-order scalar relativistic calculation (labeled, as usual, as either DKH or DK)²² with a large AE basis set called cc-pVSZ-DK.²³ This basis set is specifically optimized for relativistic calculations and is used to obtain a D_e reference value for that functional. For brevity, this reference value is labeled NCBS-DK (NCBS denotes “nearly complete basis set”). For each ECP calculation, with either a nonrelativistic (NR) or a DKH relativistic treatment of valence electrons, we calculate a deviation from the relativistic NCBS-DK value; we will call this the complete error. The complete mean unsigned error (C-MUE) is the mean unsigned deviation of the 11 calculated values for each kind of dissociation limit from their corresponding NCBS-DK reference values, and it is a measure of how well the whole treatment approaches the complete basis set limit including relativistic effects. The C-MUE is also called the basis-set incompleteness error.

Similarly, we use a nonrelativistic calculation with the cc-pVSZ basis set (which is specifically optimized for nonrelativistic calculations) to obtain a nearly complete-basis-set nonrelativistic reference value, labeled NCBS-NR, of D_e of Ti_2 for each dissociation limit and for a given functional. The mean unsigned deviation of calculations with each ECP and its corresponding valence basis set from the corresponding NCBS-NR reference values is calculated for the 11 functionals and is called the nonrelativistic mean unsigned error (NR-MUE). The NR-MUE is a measure of how well this treatment approaches the nonrelativistic complete basis set limit.

As mentioned in Introduction, the error from the density functional itself interacts with the error from the basis set; therefore, instead of comparing to experimental results, we compare the ECP results to the AE NCBS values obtained with the same functional. We do this for both relativistic and nonrelativistic ECPs.

In this way, one can largely decouple the errors that are intrinsic to a given functional, errors from the treatment of relativistic effects, and errors resulting from the choice of ECP and valence basis set. By averaging over 11 functionals we get a robust estimation of the typical errors incurred by use of a given ECP to represent the core electrons for DFT calculations because averaging over 11 functionals avoids the bias due to the choice of a particular functional.

In this first test set, we first test the five popular small-core ECPs, each used with its own valence basis set, that is, with the basis set originally proposed by the developers of that ECP. Next we test some nonstandard combinations of these ECPs with AE basis sets or with valence basis sets designed, optimized, or designated for other ECPs. The five tested ECPs include three RECPs:

- the multiconfiguration Hartree–Fock adjusted relativistic Stuttgart ECP with perturbative corrections added from Dirac-Hartree–Fock results (originally called MDF10²⁴ and also often called SDD),
 - the relativistic compact effective potential (CEP,²⁵ which is also sometimes called the Stevens-Basch-Krauss-Jasien (SBKJ) potential),
 - the CRENL RECP derived from numerical Dirac–Fock calculations;²⁶
- and two nonrelativistic ECPs (NRECPs):
- the Los Alamos²⁷ small-core NRECP,
 - the MHF10²⁴ Stuttgart NRECP.

MDF10 and MHF10 are energy-adjusted ECPs, and the others are shape-consistent ECPs. The energy-adjusted ECPs are adjusted to more than a single reference state so that they are expected to describe states with different d occupation in a more balanced way. The CEP shape-consistent ECP is also obtained in this way. In all ECP calculations in the paper itself (whether employing an NRECP or an RCP), the valence electrons are treated nonrelativistically. Two calculations presented in the Supporting Information confirm the expected small effect of including scalar relativistic effects in the treatment of noncore electrons.

For comparison, we also test the performance of popular AE def2 basis sets²⁸ def2-TZVP and def2-QZVP and the minimally augmented def2 basis sets²⁹ ma-TZVP and ma-QZVP. The relativistic def2-QZVP calculations have been recommended for accurate calculations of small arsenic species in our recent work, and the nonrelativistic def2-TZVP calculations have been found to be reasonably accurate for the properties in which relativistic effect is not significant.

For brevity, we will sometimes label the combination of an ECP and a valence basis set as an “ECP basis set”. We denote each nonstandard ECP basis set by “basis[ECP]”, where “basis” is the well-known name of the basis set and “ECP” is an abbreviation for the ECP. Lanl2 denotes the small-core Los Alamos NRECP; CEP denotes the CEP RECP; MDF denotes the MDF10 RECP; and MHF denotes the MHF10 NRECP. For the standard ECP basis sets, that is, using an ECP and its own designated basis set, or for the use of an AE basis set, we just use the well-known name for it, for example, Lanl2DZ, CEP-121G, or def2-TZVP. A suffix “-C” is added in the denotation if Cartesian d, f, or g subshells are used in calculations instead of the default spherical harmonic d, f, or g basis functions.

Based on the averages of obtained C-MUEs and NR-MUEs for D_e values of two kinds of dissociation of Ti_2 , we selected several more accurate ECP basis sets and the def2-TZVP AE basis set for further tests with a larger test set using the same 11 density functionals. The test set includes 10 cases: the two kinds of

dissociation energy of Ti_2 of the first test set, the first excitation energy E_x ($4s^23d^2 \rightarrow 4s^13d^3$) of Ti atom, and D_e values of TiV^+ ($^3\Delta$ state, dissociating to a neutral Ti atom and a V^+ cation), TiH ($^4\Phi$), Fe_2 ($^7\Delta_u$), FeO ($^5\Delta$), Sc_2 ($^5\Sigma_u^-$), ScH ($^1\Sigma^-$), and Cu_2 ($^5\Sigma_g^+$). For the cases of TiH , Fe_2 , FeO , Sc_2 , ScH , and Cu_2 , only the ground-state dissociation limit is considered. In calculations with these selected ECP basis sets and in the def2-TZVP AE calculations, the ma-TZVP AE basis set always is used for the oxygen atom of FeO , and the def2-TZVP AE basis set is used for hydrogen in the cases of ScH and TiH .

The same strategy is used for this second test set, and the performance of ECP basis sets for all 11 functionals is estimated by calculating C-MUE and NR-MUE for each case and then averaging over the ten cases. Errors averaged over the ten cases are prefixed by “A-” (which should not be confused with the “M” that denotes a mean over several density functionals). In the calculations of NCBS-DK (NCBS-NR) reference values for each functional and each case, the NCBS-DK (NCBS-NR) basis set for 3d TM metals and H atom is always cc-pV5Z-DK (cc-pV5Z),^{22,30} and the NCBS-DK (NCBS-NR) basis set for O atom is a minimally augmented cc-pV5Z-DK (cc-pV5Z) basis set (ma-cc-pV5Z-DK (ma-cc-pV5Z)) because of the negative partial charge on O. The minimally augmented basis set ma-cc-pV5Z-DK or ma-cc-pV5Z is obtained by adding a set of diffuse s and p functions to the cc-pV5Z-DK or cc-pV5Z basis set for nonhydrogenic elements, with the exponents of the most diffuse s or p functions of cc-pV5Z-DK or cc-pV5Z basis set divided by a factor of 3, as recommended previously.^{4,29}

To identify an ECP that works well for a given functional, we also calculate the mean unsigned deviation of various ECP basis sets from the NCBS-DK reference value over ten cases for each given functional. In addition, for further understanding, the use of these ECP basis sets for HF calculations and for calculations with two additional functionals SVWN³¹ and HSE³² will be also tested in this part of the investigation. Through comparison with the performance of the def2-TZVP AE basis set for each functional, the use of each special ECP basis set is discussed in detail.

All calculations are carried out using the *Gaussian 09*³³ electronic-structure package. For each molecule, the same geometry was used for all calculations because the use of any reasonable bond length is equally good for testing whether the ECP represents the effect of the core electrons. (Nevertheless, for completeness, we confirm at the beginning of section 3 that the complete mean unsigned errors in dissociation energies with a fixed reasonable bond length are in good agreement with those obtained with individually optimized bond lengths.) The experimental bond lengths are used for $R_{Ti-Ti} = 1.943 \text{ \AA}$,³⁴ $R_{Fe-Fe} = 2.02 \text{ \AA}$,³⁴ $R_{Cu-Cu} = 2.219 \text{ \AA}$,³⁴ $R_{Ti-H} = 1.779 \text{ \AA}$,³⁵ and $R_{Fe-O} = 1.616 \text{ \AA}$.³⁶ The value of R_{Sc-Sc} equal to 2.63 \AA is taken from the DFT calculation in ref 34; $R_{Sc-H} = 1.7709 \text{ \AA}$ is obtained here by the M05/NCBS-DK method; and $R_{TiV^+} = 2.4287 \text{ \AA}$ is obtained here by the M05/def2-TZVP method. For each single-point electronic structure calculation, the internal stability³⁷ of wave function has been tested. If instability is found, the wave function is reoptimized with the appropriate reduction in constraints, until it is stable.

3. RESULTS AND DISCUSSION

For convenience of the reader understanding the trends, all values of errors in the tables are rounded to the nearest 0.1 kcal/mol. For specialists, another set of tables showing the hundredths place is given in the Supporting Information.

Table 1. C-MUE of Equilibrium Bond Length R_e (Å) and C-MUE of D_e (kcal/mol) of TiH and Cu_2 over 11 Density Functionals

	TiH			Cu_2		
	R_e	D_e^a	D_e^b	R_e	D_e^a	D_e^b
NCBS-DK	0	0	0	0	0	0
NCBS-NR	0.0032	1.7	1.7	0.0303	2.7	2.9
def2-TZVP (DK)	0.0019	1.0	1.0	0.0146	0.8	0.9
def2-TZVP (NR)	0.0012	0.8	0.8	0.0445	3.7	4.0
Lan12DZ	0.0220	2.3	2.4	0.0163	1.8	1.8
ma-sc-SVP	0.0338	2.6	2.6	0.0122	13.2	13.1
MDF[Lan12]-C	0.0035	1.0	1.0	0.0316	3.8	4.0

^a D_e is calculated with individually optimized bond length. ^b D_e is calculated with fixed bond length.

Before the systematic studies, we take TiH and Cu_2 as examples to compare the errors calculated from D_e values with particularly optimized bond length R_e for a certain method and basis to those obtained with a fixed reasonable bond length. As shown in Table 1, for both molecules, the C-MUE values of the calculated D_e with the optimized bond length over 11 density functionals (M05, M06-L, M06, BLYP, ω B97X-D, τ HCTHhyb, G96LYP, mPWLYP, B3LYP, X3LYP, and BPBE) are similar to those obtained with the fixed bond length, even for some basis choices which give a large geometry deviation from the NCBS-DK limit; this indicates that the protocol used here is insensitive to choice of geometry within the near equilibrium region.

3.1. D_e of Ti_2 . Table 2 lists the calculated MUEs (NR-MUE, C-MUE) using four AE basis sets (def2-QZVP, def2-TZVP, ma-QZVP, and ma-TZVP) and 24 ECP basis sets (the details are in Table 2, and more ECP basis sets test results can be found in the Supporting Information), for D_e values of two kinds of dissociations of Ti_2 , averaged over the 11 density functionals (M05, M06-L, M06, BLYP, ω B97X-D, τ HCTHhyb, G96LYP, mPWLYP, B3LYP, X3LYP, and BPBE). The last two columns give the average of MUEs (A-NR-MUE, A-C-MUE) of the two kinds of D_e values of Ti_2 , so each value in these columns is an average over 22 data.

3.1.1. AE Basis Sets. As shown in Table 2, the nonrelativistic calculations with def2- x ZVP and ma- x ZVP ($x \geq T$) AE basis sets all have A-NR-MUE values of <0.5 kcal/mol; it is not necessary to use an ma- basis set instead of a def2- basis set on the metal in the D_e calculations of 3d transition metal–metal bond. In a similar way, the def2- basis sets tested with relativistic treatment show acceptably small errors with A-C-MUE being smaller than 0.7 kcal/mol, even though these basis sets are optimized for nonrelativistic calculations. It is noted that in our previous study⁴ for arsenic, relativistic calculations with def2-TZVP were not recommended because the def2-TZVP basis set is overpolarized by f functions for relativistic valence orbitals of arsenic. We note that def2-TZVP is as good as or slightly better for the two kinds of dissociations of Ti_2 than the def2-QZVP basis set for relativistic treatment. Using the C-MUEs of nonrelativistic calculations with cc-pV5Z, which is the NCBS-NR basis set, we can roughly estimate the relativistic effect for the two dissociation limits of Ti_2 . The ground state dissociation limit of Ti_2 is found to have a larger relativistic effect, which is about twice that of the dissociation to the first excited state of Ti atom. The average of the absolute values of relativistic effect for the two dissociation

limits of Ti_2 is ~ 3.2 kcal/mol, which should be the smallest error for AE nonrelativistic treatment for Ti_2 dissociations. In order to get more accurate results, a relativistic treatment must be used.

3.1.2. ECP Basis Sets. According to Table 2, relativistic DFT calculations with the def2-TZVP AE basis set can decrease the complete error to 0.6 kcal/mol. Can one also achieve this with an RECP? If not, can the RECP at least reduce the A-C-MUE to a value below the 3 kcal/mol achievable with NR AE calculations? We shall see that the answers are no and no. (We remind the reader that all errors in and all conclusions in this section are for errors averaged over 11 density functionals.)

Table 2 shows a large number of calculations on Ti_2 , which was studied in the most detail because it turned out to be a very difficult test case. The table shows that the performance of all the ECP basis sets tested (including both RECPs and NRECPs) is much worse than those of the AE basis sets, even worse than those of the nonrelativistic AE basis set calculations. Most standard ECP basis sets and some nonstandard ECP basis sets are singled out for having relatively better performance in Table 2. Next we discuss a few examples showing this.

The smallest A-C-MUE is 7.6 kcal/mol obtained with the MDF[CEP]-C ECP basis set, which is a nonstandard combination of CEP RECP and the valence basis set particularly developed for use with the MDF10 RECP except that “-C” denotes the Cartesian d and f functions are used in calculations. The CEP-121G standard ECP basis adopts the same CEP RECP but has larger A-C-MUE value (10.4 or 10.6 kcal/mol for spherical harmonic or Cartesian d functions) due to using a relatively smaller valence basis set. This example provides an illustration of why we considered nonstandard ECP basis sets. It shows that sometimes the error in a standard ECP basis set is not due entirely to the ECP itself but rather has a significant component due to the originally proposed valence basis set. But in this case we can still reduce the average complete error by only 26% by using a better valence basis set.

The Los Alamos small-core NRECP is also a relatively good ECP. Since it is fitted to nonrelativistic calculations, the relativistic effect is not taken account. Hence, although the Los Alamos NRECP with its own valence basis sets (Lan12DZ-C and Lan12DZ) has a worse performance than MDF[CEP]-C as compared to NCBS-DK reference values, it has the best performance compared to NCBS-NR values. We tried to improve the performance of the Los Alamos ECP by adding more polarization functions and by decontraction, as discussed next. Lan12DZ(f)-C has an additional f polarization shell for the valence basis set as compared to Lan12DZ-C; these additional f polarization functions slightly improve its performance relative to Lan12DZ-C with the average error decreasing by 0.3–0.5 kcal/mol. Lan12TZ and Lan12TZ(f) decontract the s and p functions of valence basis set of Lan12DZ, and Lan108 decontracts all s, p, and d functions; these basis sets, although further flexibilized by uncontractions, still lead to large MUEs due to the incompleteness of the basis set.

Although MDF10 and CRENBL are both RECPs fitted to relativistic calculations, they perform worse than Los Alamos and MHF10 NRECPs for the Ti_2 dissociation energies, especially the CRENBL RECP, which has an A-C-MUE of 12.5 kcal/mol. When spherical harmonic d, f, or g functions are used, MDF2fg has a smaller A-C-MUE value than that of MDF due to one additional f shell and one g shell added to valence basis set, which was recommended by Martin.³⁸ However, the use of Cartesian functions gets the opposite results: the AMUEs of MDF2fg-C are found to be larger than those of MDF-C.

The nonstandard ECP basis sets in which an ECP is simply combined with an AE basis set usually perform very badly. This is

Table 2. MUE (kcal/mol) over 11 Density Functionals for D_e of Ti_2^a

	basis function	ECP	RECP	type ^b	$Ti_2 \rightarrow 2Ti$ (³ F)		$Ti_2 \rightarrow 2Ti$ (⁵ F)		A-NR-MUE	A-C-MUE
					NR-MUE	C-MUE	NR-MUE	C-MUE		
cc-pVSZ-DK	9s,8p,6d,4f,3g,2h,1i			DK	4.3	0.0	2.1	0.0	3.2	0.0
def2-TZVP	6s,4p,4d,1f			DK	5.2	0.9	2.0	0.3	3.6	0.6
def2-QZVP	11s,6p,5d,3f,1g			DK	3.4	0.9	2.6	0.5	3.0	0.7
def2-TZVP	6s,4p,4d,1f			NR	0.6	3.7	0.3	2.4	0.5	3.0
ma-TZVP	7s,5p,4d,1f			NR	0.5	4.0	0.2	2.1	0.3	3.0
def2-QZVP	11s,6p,5d,3f,1g			NR	0.6	4.5	0.4	1.7	0.5	3.1
cc-pVSZ	9s,8p,6d,4f,3g,2h,1i			NR	0.0	4.3	0.0	2.1	0.0	3.2
ma-QZVP	12s,7p,5d,3f,1g			NR	0.4	4.7	0.3	1.7	0.4	3.2
MDF[CEP]-C	6s,5p,3d,1f	CEP	yes	NR	9.6	6.8	7.9	8.5	8.7	7.6
Lanl2DZ(f)-C	3s,3p,2d,1f	Los Alamos	no	NR	6.7	5.7	9.5	10.1	8.1	7.9
MDF[Lanl2]	6s,5p,3d,1f	Los Alamos	no	NR	8.3	5.8	10.1	10.7	9.2	8.3
MD[Lanl2]-C	6s,5p,3d,1f	Los Alamos	no	NR	9.9	7.1	8.9	9.5	9.4	8.3
Lanl2DZ-C	3s,3p,2d	Los Alamos	no	NR	7.1	6.0	10.1	10.7	8.6	8.3
MHF-C	6s,5p,3d,1f	MHF10	no	NR	10.4	7.6	10.1	10.7	10.2	9.2
def2-QZVP[Lanl2]	11s,6p,5d,3f,1g	Los Alamos	no	NR	8.1	12.4	7.9	6.0	8.0	9.2
Lanl2DZ	3s,3p,2d	Los Alamos	no	NR	5.9	8.2	10.0	10.6	8.0	9.4
m-Lanl2DZ	3s,3p,2d	Los Alamos	no	NR	6.0	8.8	9.8	10.4	7.9	9.6
MDF2fg	6s,5p,3d,2f,1g	MDF10	yes	NR	13.2	9.9	9.3	9.8	11.2	9.9
MDF-C	6s,5p,3d,1f	MDF10	yes	NR	13.6	10.2	9.2	9.8	11.4	10.0
CEP-121G[Lanl2]-C	4s,4p,3d	Los Alamos	no	NR	13.1	9.6	9.9	10.5	11.5	10.1
MDF	6s,5p,3d,1f	MDF10	yes	NR	13.4	10.0	9.6	10.1	11.5	10.1
CEP-121G	4s,4p,3d	CEP	yes	NR	13.9	10.3	9.8	10.4	11.9	10.4
CEP-121G-C	4s,4p,3d	CEP	yes	NR	14.4	10.8	9.8	10.4	12.1	10.6
MDF2fg-C	6s,5p,3d,2f,1g	MDF10	yes	NR	14.8	11.2	9.5	10.1	12.1	10.7
Lanl2TZ(f)	5s,5p,3d,1f	Los Alamos	no	NR	14.5	11.7	10.0	10.6	12.3	11.1
Lanl2TZ	5s,5p,3d	Los Alamos	no	NR	14.9	12.0	10.3	10.9	12.6	11.4
CEP-121G[MDF]-C	4s,4p,3d	MDF10	yes	NR	14.9	11.3	11.1	11.7	13.0	11.5
Lanl08	5s,5p,5d	Los Alamos	no	NR	14.9	12.1	10.4	11.0	12.7	11.6
def2-TZVP[Lanl2]	6s,4p,4d,1f	Los Alamos	no	NR	12.8	17.1	5.5	6.1	9.1	11.6
CRENBL	7s,6p,6d	CRENBL	yes	NR	17.2	13.6	10.8	11.4	14.0	12.5
Lanl2MB	2s,2p,1d	Los Alamos	no	NR	10.8	9.8	19.1	20.1	15.0	15.0
ma-sc-SVP	5s,4p,2d,2f	MDF10	yes	NR	5.8	5.9	26.7	24.6	16.2	15.3

^aThe well-known names of AE basis sets and standard combinations of ECP and valence basis sets are used. The nonstandard combinations of ECP and basis set are denoted as “basis[ECP]”, where “basis” is the well-known name of the basis set and “ECP” is an abbreviation for the name of ECP that is used. Lanl2 denotes small-core (10e) Los Alamos NRECP; CEP denotes CEP RECP; MDF denotes MDF10 RECP; MHF denotes MHF10 NRECP. If Cartesian d, f, or g functions are used in calculations, a suffix “-C” is used. ^bThis column refers to how the valence electrons are treated; some calculations in which they are treated relativistically are presented in the Supporting Information.

attributed to these AE basis sets being incomplete and not flexible enough. The pseudovalence orbitals in ECP calculations are different from the valence orbitals in AE calculations in shape and size extent. The exponents of basis functions and the contraction coefficients from an incomplete AE bases sets are therefore not suitable for description of pseudovalence orbitals. Relatively good results for combinations of this type are only observed for def2-QZVP[Lanl2], where the relatively complete AE basis set def2-QZVP is used as the valence basis set.

3.2. D_e of FeO, Cu₂, ScH, TiH, Sc₂, Fe₂, and TiV⁺ and E_x of Ti. Based on A-C-MUE values in Table 2 for Ti_2 dissociation energies, we selected several ECP basis sets with relatively good performance averaged over 11 density functionals for further tests. The criterion is that A-C-MUE in Table 2 is smaller than 10.0 kcal/mol. In addition, the MDF2fg-C and ma-sc-SVP with relatively larger A-C-MUE values are also selected for special

interest (which will be specified in the later discussion). However the def2-QZVP[Lanl2] and m-Lanl2DZ choices are excluded although their A-C-MUEs are smaller than 10.0 kcal/mol. The def2-QZVP[Lanl2] is excluded because it has the same CPU time as def2-QZVP AE calculations but behaves worse due to the use of an ECP. So it is not very attractive to further test this ECP basis set. The m-Lanl2DZ is a slightly modified Lanl2DZ ECP basis set, and its valence basis set has one more p primitive function and different p contracted functions compared with Lanl2DZ. This small change of valence basis sets affects the performance only very slightly. Therefore, we only choose Lanl2DZ for later tests. Thus, in the rest of this section, we employ 11 ECP basis sets (MDF[CEP]-C, Lanl2DZ(f)-C, MDF[Lanl2], MDF[Lanl2]-C, Lanl2DZ-C, MHF-C, Lanl2DZ, MDF2fg, MDF-C, MDF2fg-C, ma-sc-SVP) to perform DFT calculations with the same 11 density functionals as used in section 3.1, but here we apply them to D_e of

Table 3. NR-MUE (kcal/mol) for D_e of Diatomic Molecules and the First Excitation Energy E_x of Ti Atom Using Different ECP and Basis Set over 11 Functionals

	D_e of diatomic molecules									E_x of Ti	A-NR-MUE(10)
	FeO ^a	Cu ₂	ScH ^b	TiH ^b	Sc ₂	Fe ₂	Ti ₂ ^c	Ti ₂ ^d	TiV ⁺		
NCBS-NR ^e	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
def2-TZVP (DK) ^f	2.6	2.0	0.6	2.7	1.5	6.0	5.2	2.0	1.3	3.6	2.7
def2-TZVP (NR) ^g	1.6	1.1	0.6	1.0	0.4	1.0	0.6	0.3	0.8	0.2	0.8
Lan12DZ-C	4.0	3.5	1.0	1.2	1.7	5.6	7.1	10.1	4.3	1.5	4.0
Lan12DZ(f)-C	4.2	3.6	1.0	1.5	2.1	6.5	6.7	9.5	4.3	1.4	4.1
Lan12DZ	4.2	4.2	0.7	1.5	4.1	6.6	5.9	10.0	4.0	3.0	4.4
MDF[Lan12]	3.7	2.4	0.8	1.5	1.7	8.4	8.3	10.1	6.6	1.2	4.5
MDF[CEP]-C	7.1	0.5	1.0	2.6	2.3	9.5	9.6	7.9	4.1	2.5	4.7
MDF[Lan12]-C	5.3	1.1	0.9	2.6	2.1	9.4	9.9	8.9	5.2	2.7	4.8
MHF-C	7.0	1.8	1.3	2.9	3.9	9.3	10.4	10.1	6.7	2.2	5.6
MDF2fg	6.9	1.2	1.7	4.7	5.4	13.4	13.2	9.3	5.4	5.1	6.6
MDF-C	8.0	1.2	1.5	4.6	5.4	14.5	13.6	9.2	5.9	5.5	6.9
MDF2fg-C	6.9	1.8	1.5	4.7	5.8	16.1	14.8	9.5	5.3	6.1	7.2
ma-sc-SVP	3.1	16.1	1.1	4.4	3.2	13.9	5.8	26.7	2.6	12.8	8.9

^aIn the calculations with ECP for metals, the basis set for O is ma-TZVP. ^bIn the calculations with ECP for metals, the basis set for H is def2-TZVP. ^cTi₂ → 2Ti (3F). ^dTi₂ → 2Ti (5F). ^eNCBS-NR: nonrelativistic results with NCBS-NR basis set. NCBS-NR basis set is cc-pV5Z for all transition metals and H atom, ma-cc-pV5Z for O atom. ^fRelativistic calculations with the def2-TZVP basis set. ^gNonrelativistic calculations with the def2-TZVP basis set.

Table 4. C-MUE (kcal/mol) for D_e of Diatomic Molecules and the First Excitation Energy E_x of Ti Atom Using Different ECP and Basis Set over 11 Functionals

	D_e of diatomic molecules									E_x of Ti	A-C-MUE(10)
	FeO ^a	Cu ₂	ScH ^b	TiH ^b	Sc ₂	Fe ₂	Ti ₂ ^c	Ti ₂ ^d	TiV ⁺		
NCBS-DK ^e	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
NCBS-NR ^f	1.4	2.9	0.2	1.7	1.4	5.1	4.3	2.1	1.0	3.2	2.3
def2-TZVP (DK) ^g	1.8	0.9	0.5	1.0	0.2	1.5	0.9	0.3	0.8	0.4	0.8
def2-TZVP (NR) ^h	1.6	4.0	0.6	0.8	1.5	4.3	3.7	2.4	1.1	3.0	2.3
MDF[CEP]-C	5.7	2.7	0.9	1.2	1.2	7.1	6.8	8.5	4.7	1.4	4.0
MDF[Lan12]-C	3.9	4.0	0.9	1.0	1.3	7.1	7.1	9.5	5.8	1.4	4.2
Lan12DZ(f)-C	3.3	1.5	1.0	1.4	2.3	8.7	5.7	10.1	4.8	3.2	4.2
Lan12DZ-C	3.0	1.5	0.9	1.5	2.1	8.5	6.0	10.7	4.7	3.6	4.2
MDF[Lan12]	3.2	5.3	0.7	1.3	1.7	8.7	5.8	10.7	7.2	3.5	4.8
MHF-C	5.6	4.7	1.3	1.2	2.5	7.4	7.6	10.7	7.4	1.9	5.0
MDF2fg	5.5	1.7	1.7	2.9	4.0	9.1	9.9	9.8	6.1	1.9	5.3
Lan12DZ	3.4	1.8	0.8	2.4	5.4	10.4	8.2	10.6	4.4	5.5	5.3
MDF-C	6.6	1.7	1.5	2.9	4.0	9.9	10.2	9.8	6.5	2.3	5.4
MDF2fg-C	5.5	1.1	1.5	3.0	4.4	11.2	11.2	10.1	6.0	2.9	5.7
ma-sc-SVP	2.4	13.1	1.0	2.6	4.3	8.9	5.9	24.6	2.7	9.6	7.5

^aIn the calculations with ECP for metals, the basis set for O is ma-TZVP. ^bIn the calculations with ECP for metals, the basis set for H is def2-TZVP. ^cTi₂ → 2Ti (3F). ^dTi₂ → 2Ti (5F). ^eRelativistic results with the NCBS-DK basis set. The NCBS-DK basis set is cc-pV5Z-DK for all transition metals and H atom, ma-cc-pV5Z-DK for O atom. ^fNonrelativistic results with the NCBS-NR basis set. The NCBS-NR basis set is cc-pV5Z for all transition metals and H atom, ma-cc-pV5Z for O atom. ^gRelativistic calculations with the def2-TZVP basis set. ^hNonrelativistic calculations with the def2-TZVP basis set.

FeO, Cu₂, ScH, TiH, Sc₂, Fe₂, and TiV⁺ and E_x of Ti. For comparison, relativistic and nonrelativistic calculations with the def2-TZVP AE basis set are also carried out. All the NR-MUEs and C-MUEs of 11 ECP basis sets and of def2-TZVP AE calculations are given respectively in Tables 3 and 4, which also includes those for two kinds of dissociations of Ti₂ from section 3.1. For each basis, the average of MUEs over the ten cases (A-NR-MUE(10) and A-C-MUE(10)) are calculated and shown in the last columns of

Tables 3 and 4, so that each value in the final columns of these tables is an average over 110 data.

3.2.1. Comparison with Nonrelativistic NCBS Results. According to the NR-MUEs shown in Table 3, the Los Alamos NRECP is the best ECP. The valence basis set normally designated for use with the Los Alamos NRECP performs slightly better when Cartesian d functions are used in calculations (Lan12DZ-C has a 0.4 kcal/mol smaller A-NR-MUE(10) than Lan12DZ).

By comparison of the performance of Lanl2DZ(f)-C and Lanl2DZ-C, the additional f polarization functions in valence basis set are found to have less effect on the present investigations. MDF-[Lanl2], which is a combination of the valence basis set originally developed for the MDF10 RECP and the Los Alamos NRECP, works better when using spherical harmonic d and f functions. Lanl2DZ and MDF[Lanl2] are observed to have similar values of A-NR-MUE(10). Thus, when the Los Alamos NRECP is used, Lanl2DZ-C and MDF[Lanl2]-C have the smallest and largest A-NR-MUE(10) values respectively. Altogether though, the performance of all the ECPs is poor in this test. In particular, if relativistic effects are not taken into account, the use of ECP basis sets has a 5–12 times larger A-NR-MUE(10) than that of def2-TZVP nonrelativistic AE calculations.

The relativistic effects for 3d TM metals are usually not negligible, and they must be taken into account. For the present tests, the relativistic effect is estimated to be up to 5 kcal/mol. Hence, it is more important to check the performance of the 11 ECP basis sets and def2-TZVP AE basis set relative to NCBS-DK results, i.e. to check their C-MUEs. We do this in section 3.2.2.

3.2.2. Comparison with Relativistic NCBS Results. Comparison of Tables 3 and 4 shows that the A-C-MUE(10) of def2-TZVP relativistic calculations is similar to the A-NR-MUE(10) of def2-TZVP nonrelativistic calculations. This indicates that relativistic calculations with the def2-TZVP basis set optimized for nonrelativistic calculations are feasible for DFT calculations of 3d TM metals, in spite of their relatively bad performance for arsenic. Therefore, in succeeding discussions, we will compare the C-MUEs of ECP basis sets with those of def2-TZVP relativistic calculations.

Again, as seen in Table 4, the CEP RECP and Los Alamos NRECP are the two best ECPs for the present DFT study as compared to NCBS-DK calculations. Here, because the comparison is now to reference values including relativistic effects, the relativistic CEP performs better than the Los Alamos NRECP. It is surprising that MDF-C using the relativistic MDF10 ECP has a larger A-C-MUE(10) than MHF-C employing the nonrelativistic MHF10 ECP. Overall though, the performance of the ECPs is disappointing. In particular, Table 4 shows that A-C-MUE(10) values of calculations employing ECP basis sets are 5–9 times larger than the A-C-MUE(10) of def2-TZVP relativistic AE calculations.

The ma-sc-SVP basis set is an ma-SVP basis set specially modified for combination as a valence basis set with the MDF10 RECP, and in our recent work,⁴ this ECP basis set was found to perform relatively well for arsenic; therefore, it was recommended for large TM-containing arsenic systems because it is available for all elements up to radon. We also mentioned that one must be careful about using it because it had not been tested systematically for 3d TM elements. According to the present investigation, ma-sc-SVP is not a stable enough ECP basis set; although it does perform relatively well for some cases, such as D_e of FeO, ScH, TiV^+ , and ground state dissociation of Ti_2 , it performs very badly for D_e of Cu_2 , for the excited dissociation limit of Ti_2 , and for excitation of Ti. This leads to the result that ma-sc-SVP has the worst performance of all 11 tested ECP basis sets in this section, and it shows why broad testing is required to draw reliable conclusions. We note that the MDF10³⁹ ECP used in the ma-sc-SVP basis set for As is a multiconfiguration Dirac-Hartree-Fock adjusted fully relativistic ECP.

3.2.2.1. D_e of Diatomic Molecules. All ECP basis sets for D_e calculations of transition metal-metal bonds of Fe_2 , Ti_2 , and

TiV^+ have very bad performance as compared to NCBS-DK reference calculations that include the relativistic effect. However, ECP calculations work for Cu_2 and Sc_2 . For Cu_2 , although they still perform worse than def2-TZVP relativistic AE calculations, most ECP basis sets, except for MDF[Lanl2], MHF-C, and ma-sc-SVP, give much better results than def2-TZVP nonrelativistic AE calculations. MDF2fg-C is especially good for D_e of Cu_2 , and it almost behaves as well as def2-TZVP relativistic calculations. Its C-MUE value is only 1.1 kcal/mol, much smaller than the error (2.9 kcal/mol) introduced by nonrelativistic NCBS calculations. The relatively good performance of ECPs for Cu_2 can be attributed to the fully occupied d shell of Cu, which avoids the variable s-d occupancy problem mentioned in the second paragraph of section 1. The ECP calculations with MDF[CEP]-C and MDF[Lanl2]-C work well for D_e of Sc_2 , for which their C-MUEs are about 1.2 kcal/mol. Other ECP basis sets also perform better for Sc_2 than for Fe_2 , Ti_2 , and TiV^+ . We note that the valence electron configuration of Sc_2 ($^5\Sigma_u$) is $[\sigma_g^2 \pi_u^1 \pi_u^1 \sigma_g^1 \sigma_u^1]$, while the valence electron configurations of Fe_2 ($^7\Delta_u$), Ti_2 ($^3\Delta_g$), and TiV^+ ($^3\Delta$) are respectively $[\sigma_g^2 \pi_u^2 \pi_u^2 \sigma_g^2 \delta_g^2 \delta_g^1 \delta_u^1 \delta_u^1 \pi_g^1 \pi_g^1 \sigma_u^1]$, $[\sigma_g^2 \pi_u^2 \pi_u^2 \sigma_g^1 \delta_g^1]$, and $[\sigma^2 \pi^2 \pi^2 \sigma^1 \delta^1]$. In the latter three cases, two degenerate δ orbitals have different electron occupations. This implies larger multireference character of these bonds. The accurate description of multireference character with DFT methods can place a greater demand on the transferability of ECPs, as discussed in the fourth paragraph of section 1.

The calculations employing suitable ECP basis sets for ScH and TiH can obtain very similar results to def2-TZVP relativistic calculations. This may result because metal-hydrogen bonding causes less changes the electronic environment of the transition metal atom than does bonding to other elements.

3.2.2.2. E_x of Ti. To some extent, the C-MUEs for E_x of Ti can represent the transferability of ECPs for different s-d occupancies. C-MUE values of MDF[CEP]-C and MDF[Lanl2]-C are 1.4 kcal/mol for E_x of Ti. This is a 7.5% error since the experimental E_x of Ti is only 18.68 kcal/mol.⁴⁰ The ma-sc-SVP ECP basis set performs very badly for E_x of Ti, which explains its large C-MUE for Ti_2 dissociation to excited states of Ti atom. The ECP basis sets which perform well for E_x of Ti gives relatively better results for diabatic dissociation of Ti_2 than for ground dissociation.

3.3. Specific Accuracies of ECP Basis Sets for Particular Density Functionals. According to the above results, in the four ECPs tested for all ten cases, the CEP RECP and the Los Alamos NRECP have been found to be relatively better ECPs, and MDF10 and MHF10 to be relatively worse for DFT calculations of 3d TM metals when the results are averaged over 11 density functionals. However, we do expect that specific ECP basis sets can be better for certain specific functionals. Our goal in this subsection is to look for an ECP basis set that, although it is not the best one when averaged over a diverse set of DFT methods, works well for a given functional. Therefore, for each of the 11 functionals in the present study so far (M05, M06-L, M06, BLYP, ω B97X-D, τ HCTHhyb, G96LYP, mPWLYP, B3LYP, X3LYP, and BPBE) and for two additional functionals (SVWN and HSE), we also calculate the average complete unsigned deviation (A-C-UE) for the 11 most promising ECP basis sets and for the def2-TZVP basis set from the NCBS-DK reference values over the ten cases tested in previous sections. For comparison, the performance of these ECP basis sets for the HF method is also investigated in the same way, which is particularly interesting

Table 5. A-C-UEs (kcal/mol) over All 10 Cases for Each Density Functional and for HF

	SVWN	BLYP	G96LYP	mPWLYP	BPBE	τ HCTHhyb	B3LYP	X3LYP	ω B97X-D	HSE	M06-L	M05	M06	HF
<i>X</i>	0	0	0	0	0	15	20	21.8	22.2–100	25–0	0	28	27	
NCBS-DK ^a	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
NCBS-NR ^b	1.9	1.8	1.6	1.9	1.5	2.7	2.7	2.8	2.9	2.8	2.5	2.7	2.6	4.9
def2-TZVP (DK) ^c	1.1	0.7	0.7	0.6	0.8	1.1	0.7	0.7	1.1	0.9	1.0	1.4	0.4	2.9
def2-TZVP (NR) ^d	1.9	1.9	1.8	1.9	1.6	2.3	2.7	2.5	2.8	2.4	2.4	2.7	2.7	5.8
MDF[CEP]-C	1.9	4.1	4.6	4.0	3.5	3.7	2.1	2.0	2.4	1.7	3.6	6.6	7.4	4.2
MDF[Lan12]-C	2.5	4.4	5.1	4.2	4.1	4.0	2.4	2.2	2.8	2.0	4.3	5.9	6.5	4.4
Lan12DZ(f)-C	1.9	3.1	4.0	3.0	3.3	3.4	2.6	2.7	2.6	3.1	4.5	7.5	9.6	10.2
Lan12DZ-C	2.0	3.4	3.0	3.3	3.5	3.6	2.9	3.0	3.1	3.3	4.0	7.5	9.4	10.8
MDF[Lan12]	3.3	4.9	5.6	4.7	5.0	4.1	3.0	4.4	3.2	3.3	3.8	6.6	7.6	5.8
MHF-C	3.2	5.4	6.2	5.1	5.4	4.8	3.1	3.0	3.8	2.8	5.3	6.3	7.0	6.7
MDF2fg	3.3	5.7	6.3	5.5	5.0	6.0	4.1	3.9	4.6	3.3	5.6	5.6	5.8	2.4
Lan12DZ	2.9	4.2	4.3	4.1	4.1	3.5	4.7	4.9	4.8	4.5	4.8	7.6	11.3	9.0
MDF-C	3.7	5.9	6.4	5.8	5.0	6.3	4.4	4.2	5.1	3.5	6.1	5.7	6.2	2.2
MDF2fg-C	3.4	6.3	6.7	6.2	5.2	6.6	4.8	4.6	5.5	3.7	5.8	5.6	5.3	1.0
ma-sc-SVP	7.5	6.5	6.4	6.4	6.7	6.8	5.8	5.7	6.5	6.1	9.6	10.2	11.9	7.2

^a All-electron relativistic results with the NCBS-DK basis set, which is cc-pVSZ-DK for all transition metals and the H atom and ma-cc-pVSZ-DK for the O atom. ^b All-electron nonrelativistic results with the NCBS-NR basis set, which is cc-pVSZ for all transition metals and the H atom and ma-cc-pVSZ for the O atom. ^c All-electron relativistic calculations with the def2-TZVP basis set. ^d All-electron nonrelativistic calculations with the def2-TZVP basis set.

since the HF method does not require NLCC, as discussed in section 1. The calculated A-C-UEs for each functional and HF method are shown in Table 5.

We will arrange the discussion of density functionals into three groups. Subsection 3.3.1 considers the local spin density approximation (SVWN) and four local generalized gradient approximations (GGAs, in particular BLYP, G96LYP, mPWLYP, and BPBE). Subsection 3.3.2 considers two global-hybrid GGAs (B3LYP and X3LYP) and two range-separated hybrid GGAs (ω B97X-D and HSE). Subsection 3.3.3 considers one meta-GGA (M06-L) and three hybrid meta-GGAs (τ HCTHhyb, M05, and M06).

The percentage *X* of Hartree–Fock exchange in each functional is shown in the first row of Table 5. Note that for range-separated functionals, *X* depends on interelectronic distance r_{12} , and for these functionals it is shown as a range, with the first number being *X* at small r_{12} and the second number being *X* at large r_{12} .

3.3.1. SVWN, HF, and GGAs. Table 5 shows that the local spin density approximation SVWN to the exchange-correlation functional is less sensitive to the choice of ECP basis sets than are any of the other functionals tested. One main obstacle to the reliable use of ECPs in DFT studies is that ECPs convert small-*s* core regions into large-*s* core regions, where *s* is the reduced density gradient. Most functionals depend on *s*, with the local spin-density approximation being the only significant exception. Since we find that ECPs perform much better for the local spin-density approximation, we conclude that the *s* issue is one of the dominant error sources of DFT calculations employing ECPs. Since the local-spin density approximation is independent of *s*, it is expected, on that basis, to have the same ECP requirements as the HF method. However the last column of Table 5 shows that, except for the MDF10 RECP, the ECPs tested here, even though they were derived for HF methods, have even larger basis-set incompleteness error for HF than for most density functionals. However, we must recall that here we compare our ECP basis set results to the nearly complete basis set limit including relativistic

effects. The HF method significantly overestimates relativistic effects relative to DFT methods, as evidenced by their larger A-C-UE values for cc-pVSZ nonrelativistic (NCBS-NR) calculations. This explains the much worse performance of Los Alamos and MHF10 NRECPs, which are fitted to nonrelativistic AE results, when compared to that of the MDF10 RECP which is fitted to relativistic AE results. In addition, as shown in Table 5, the A-C-UEs of def2-TZVP AE calculations using the HF method are more than twice those of DFT methods, which implies slower basis set convergence in HF. Hence, due to using larger valence basis sets, the HF calculations of MDF[Lan12]-C and MDF[Lan12] perform much better than those using Lan12DZ(f)-C, Lan12DZ-C, and Lan12DZ. These complications of the HF results make it hard to compare their A-C-UEs in Table 5 directly with those of DFT.

Therefore, we instead compare the results of the local spin density approximation to those of GGAs, where the energy density definitely depends on the reduced density gradient *s*. Although the def2-TZVP calculations with SVWN and GGAs have similar A-C-UEs, GGA calculations using ECP basis sets are observed to have relatively larger basis-set incompleteness errors. As mentioned above, this shows that the *s* issue is significant. It is also noted that Lan12DZ(f)-C and Lan12DZ-C ECP basis sets are preferred for both GGAs and SVWN. MDF[CEP]-C is also reasonable for SVWN and GGAs.

3.3.2. Hybrid GGAs. Hybrid GGAs functionals (B3LYP, X3LYP, ω B97X-D, and HSE) have worse ECP performance than the local spin density approximation but better performance than GGAs. This must be due to the component of HF exchange. The DFT exchange and correlation parts of these hybrid functionals can yield reasonably accurate estimates of relativistic effects and ensure fast basis set convergence, while the portion of HF exchange that replaces a portion of GGA exchange reduces the *s* difficulty of the GGA functionals. MDF[CEP]-C works best for all hybrid GGAs, and it has a smaller A-C-UE value than def2-TZVP nonrelativistic AE calculations.

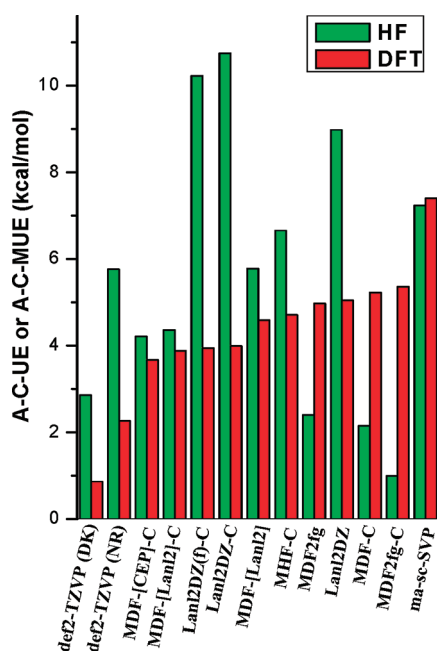


Figure 1. Average complete unsigned error (A-C-UE in kcal/mol) of HF and average complete mean unsigned error (A-C-MUE in kcal/mol) of DFT (averaged over 13 functionals) for ten cases, where the “errors” are deviations as compared to the corresponding NCBS-DK reference values.

The ω B97X-D and HSE range-separated hybrid functionals use different amounts of HF exchange for different interelectronic separation ranges. The ω B97X-D uses full HF exchange at large electron–electron distances and uses a small fraction of 22% HF exchange at short-range. HSE is a screened-Coulomb range-separated hybrid GGA, and it uses pure DF exchange (0% HF exchange) at large interelectronic separation and 25% HF exchange at short-range. Since the most important electron–electron interactions for core electrons that are described by an ECP are short-range interactions, although the two tested range-separated hybrid functionals have very different HF exchange percentages at long-range, they both behave like the two global hybrid functionals.

3.3.3. Meta-GGA Functionals. Meta-GGA functionals not only depend on the up and down spin densities and reduced density gradients but also on the up and down spin kinetic energy densities. The dependence on the kinetic energy densities makes them behave differently with regard to using ECPs. The M06-L meta-GGA is a local functional, but it behaves more like hybrid GGAs than like local GGAs. The hybrid meta-GGAs calculations (M05 and M06) with ECPs are found to have very large basis-set incompleteness error. The MDF10 RECP, which has the worst performance for both local and hybrid functionals, works best for the M05 and M06 functionals. This could be partially due to larger HF exchange. It is noted in this regard that the τ HCTHhyb hybrid meta-GGA functional with smaller HF exchange behaves more like local GGAs.

3.4. Comparison of HF and DFT for the Use of ECPs. Figure 1 is the plot of A-C-UE of HF method and A-C-MUE of DFT for all ten test cases. The A-C-MUEs of DFT methods are averaged for all 13 functionals tested in present study. As shown in Figure 1, except for the MDF10 RECP, the HF-derived ECPs have smaller basis-set incompleteness errors in DFT calculations than in HF calculations, which is surprising in light of

the theoretical underpinning of the need for NLCCs in DFT (no NLCCs are employed in the present work). As discussed in section 3.3.1, the intrinsic error of the HF method seems to dominate the complete error. Therefore, from the present investigation, it is hard to tell how important the NLCC correction is in DFT calculations.

The smallest complete mean unsigned error introduced for DFT calculations is about 3.7 kcal/mol with an appropriate ECP basis set, namely MDF[CEP]-C. Accuracies nearly as good are attained with MDF[Lan12]-C, 3.9 kcal/mol, Lan12DZ(f)-C, 3.9 kcal/mol, and Lan12DZ-C, 4.0 kcal/mol. However, none of these values is as good as an all-electron def2-TZVP relativistic (DK) calculation, 0.9 kcal/mol, or even an all-electron def2-TZVP nonrelativistic calculation, 2.3 kcal/mol. Since the results in Figure 1 are in a sense the culmination summary of the paper, they are presented in tabular form in the Supporting Information (Table S6).

4. SUMMARY

The present work systematically investigates the performance of ECPs and ECP basis sets in Hartree–Fock and DFT calculations of 3d transition metal species, where an ECP basis set is defined as a valence and subvalence basis set plus an effective core potential for a small core (for 3d transition metals, a small core is the innermost ten electrons). The investigation starts from calculations of two kinds of dissociation energy for Ti_2 using 11 common density functionals (M05, M06-L, M06, BLYP, ω B97X-D, τ HCTHhyb, G96LYP, mpWLYP, B3LYP, X3LYP, and BPBE). The performance of some Hartree–Fock-derived ECP basis sets (including three relativistic ECPs and two non-relativistic ECPs) and four popular all-electron basis sets with relativistic or nonrelativistic treatments is evaluated by comparing their predictions to what is obtained with relativistic NCBS-DK calculations and with nonrelativistic NCBS-NR calculations, where NCBS denotes the nearly complete basis set limit, where DK denotes a relativistic calculation, and NR denotes a non-relativistic one.

Based on their performance and on some special considerations, eleven ECP basis sets and the def2-TZVP all-electron basis set are chosen for further tests on the equilibrium bond dissociation energy (D_e) of FeO, Cu_2 , ScH, TiH, Sc₂, Fe₂, and TiV^+ and the electronic excitation energy of Ti, using the same 11 functionals. The eleven ECP basis sets chosen for this study are MDF[CEP]-C, Lan12DZ(f)-C, MDF[Lan12], MDF[Lan12]-C, Lan12DZ-C, MHF-C, Lan12DZ, MDF2fg, MDF-C, MDF2fg-C, and ma-sc-SVP. Both relativistic and nonrelativistic calculations are carried out with the def2-TZVP basis set. ECP calculations (except for a couple of tests in the Supporting Information) treat the valence electrons nonrelativistically. The general performance of eleven ECP basis sets and the def2-TZVP all-electron DFT calculations is evaluated based primarily on the average complete mean unsigned deviations (A-C-MUE) from the NCBS-DK all-electron relativistic results, averaged over ten cases including D_e of two kinds of dissociation limits of Ti_2 and over 11 functionals at first and then over 13 functionals. The numbers mentioned below refer to the average over 13 functionals.

The relativistic all-electron def2-TZVP DFT calculations for 3d transition metal species are found to be close to the basis-set limit and to have an A-C-MUE error value of only 0.9 kcal/mol. This result is qualitatively different from what we found in our previous study⁴ of arsenic DFT calculations, where relativistic

calculations with the def2-TZVP basis set for arsenic were not recommended. The use of ECP basis sets in DFT calculations for 3d transition metal species gives much worse results than def2-TZVP relativistic all-electron calculations for the density functionals employed in the various tests. The CEP relativistic ECP and the Los Alamos nonrelativistic ECP are, on average, the best general ECP choices for DFT calculations, and, in particular, the MDF[CEP]-C ECP basis set, which is a combination of the CEP relativistic ECP and valence basis sets developed for the MDF10 relativistic ECP, has the smallest average mean complete unsigned error (A-C-MUE), 3.7 kcal/mol, where “complete error” denotes the deviation from the nearly complete-basis-set relativistic calculations for a given functional, “mean” denotes averaging over the 13 functionals, and “average” denotes averaging over the ten cases.

The main goals in using relativistic ECPs are to decrease the size of the basis sets and to include relativistic effects in calculations in which the explicitly included electrons are still treated nonrelativistically. The second goal cannot be considered to be achieved satisfactorily since we find that results obtained using relativistic ECPs are farther from the nearly complete-basis-set relativistic limit than are polarized triple- ζ nonrelativistic all-electron calculations. Along the same lines, it is also disappointing that when DFT calculations are compared to nearly complete-basis-set relativistic results with the same density functional, the MDF10 relativistic ECP fitted to relativistic calculations has a larger A-C-MUE than does the MHF10 nonrelativistic ECP fitted to nonrelativistic calculations.

We also examined the question of whether some ECPs perform better for certain density functionals than their performance averaged over a set of diverse density functional calculations. To examine this, for each of the 11 selected common density functionals and for two additional density functionals (SVWN and HSE), the average complete unsigned deviation (A-C-UE) from the NCBS-DK reference values has been calculated for the eleven ECP basis sets and for the def2-TZVP all-electron basis set for all ten cases. For comparison, Hartree–Fock calculations are performed in the same way. In these tests, the local spin density approximation functional SVWN has the smallest A-C-UE error value, which we interpret as being due to its independence of the reduced density gradient s . GGAs (in which the energy density depends on s) have relatively worse performance than SVWN for the use of Hartree–Fock-derived ECPs. Both LSDA and GGAs are local functionals, and they work best with the Lanl2DZ(f)-C and Lanl2DZ-C ECP basis sets. The Hartree–Fock exchange used in hybrid GGAs partially avoids the difficulty caused by the s dependence of GGAs in ECP calculations, so that, even though hybrid GGAs are often not recommended functionals for transition metal chemistry, the use of ECPs with this kind of density functional introduces less basis-set incompleteness error than with GGAs. The calculations with hybrid functionals using appropriate ECP basis sets often show better basis set convergence than def2-TZVP nonrelativistic calculations. MDF-CEP-C and MDF[Lanl2]-C are the preferred ECP basis sets for hybrid functionals. The average ECP basis-set-incompleteness error for hybrid functionals calculations with the best performing ECP basis set is ~ 2 kcal/mol.

The meta-GGA density functionals, which depend not only on the up and down spin density and s but also on the up and down spin kinetic energy density, show different behaviors. The M06-L local meta-GGA behaves more like hybrid GGAs, the τ HCTHhyb meta-hybrid GGA behaves more like local functionals as far as its

compatibility with ECPs, and the M05 and M06 hybrid meta-GGAs have the largest errors introduced by ECPs. M05 and M06 with higher HF exchange work best with the MDF10 RECP as does the Hartree–Fock method.

The present investigations show that different functionals have different needs for ECP basis sets. Except for using hybrid functionals and the local spin density approximation, the use of ECP basis sets for 3d transition metals introduces an error of at least 3 kcal/mol for DFT calculations compared to the nearly complete basis set limit including the relativistic effect. Although hybrid functionals are not recommended for transition metal species with high multireference character, they work better with ECPs than do local functionals and meta-GGA functionals.

DFT calculations with an ill-suited ECP basis set can lead to a basis set error larger than 10 kcal/mol. Great caution is urged when using Hartree–Fock-derived ECPs in either HF or DFT studies of 3d transition metal systems. In order to get better results, relativistic or even nonrelativistic DFT calculations with the def2-TZVP all-electron basis set are recommended. It is emphasized again that the conclusions are for 3d transition metals and that they are obtained based on comparisons to the HF and DFT complete-basis-set limit including relativistic effects rather than by comparison to experimental data because comparison to experiment makes it hard to disentangle basis-set incompleteness from the quality of the Hartree–Fock approximation or the approximate density functional.

■ ASSOCIATED CONTENT

S Supporting Information. (1) Another version of Tables 1–5 showing the deviations rounded in the hundredths place (0.01 kcal/mol) rather than the tenths place; (2) some additional combinations of ECP and basis set for Ti_2 calculations and more details for the specialist and readers interested in this kind of detail; (3) NCBS-DK values for each case and each functional investigated in the present work and for Hartree–Fock calculations. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: truhlar@umn.edu.

■ ACKNOWLEDGMENT

This work was supported in part by the National Science Foundation by grant no. CHE09-56776 and by the Air Force Office of Scientific Research.

■ REFERENCES

- (1) Niu, S.; Hall, M. B. *Chem. Rev.* **2000**, *100*, 353.
- (2) (a) Louie, S. G.; Froyen, S.; Cohen, M. L. *Phys. Rev. B* **1982**, *26*, 1738. (b) Juan, Y. M.; Kaxiras, E. *Phys. Rev. B* **1993**, *48*, 14944. (c) Juan, Y. M.; Kaxiras, E.; Gordon, R. G. *Phys. Rev. B* **1995**, *51*, 9521. (d) Fuchs, M.; Bockstedte, M.; Pehlke, E.; Scheffler, M. *Phys. Rev. B* **1998**, *57*, 2134. (e) Porezag, D.; Pederson, M. R.; Liu, A. Y. *Phys. Rev. B* **1999**, *60*, 14132.
- (3) (a) Russo, T. V.; Martin, R. L.; Hay, P. J. *J. Phys. Chem.* **1995**, *99*, 17085. (b) van Wüllen, C. *Int. J. Quantum Chem.* **1996**, *58*, 147. (c) Han, Y.-K.; Hirao, K. *Chem. Phys. Lett.* **2000**, *324*, 453. (d) Yang, Y.; Weaver, M. N.; Merz, K. M., Jr. *J. Phys. Chem. A* **2009**, *113*, 9843.
- (4) Xu, X.; Truhlar, D. G. *J. Chem. Theory Comput.* **2011**, *7*, 2766.

- (5) (a) Raab, J.; Roos, B. O. *Adv. Quantum Chem.* **2005**, *48*, 421. (b) Buchachenko, A. A. *Chem. Phys. Lett.* **2008**, *459*, 73. (c) Harvey, J. N. *Annu. Rep. Prog. Chem. Sect. C: Phys. Chem.* **2006**, *102*, 203. (d) Cramer, C. J.; Truhlar, D. G. *Phys. Chem. Chem. Phys.* **2009**, *11*, 10.
- (6) (a) Schultz, N. E.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 4388. (b) Schultz, N. E.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 11127. (c) Furche, F.; Perdew, J. P. *J. Chem. Phys.* **2006**, *124*, 044103. (d) Riley, K. E.; Merz, K. M., Jr. *J. Phys. Chem. A* **2007**, *111*, 6044. (e) Riley, K. E.; Holt, B. T. O.; Merz, K. M., Jr. *J. Chem. Theory Comput.* **2007**, *3*, 407. (f) Tekarli, S. M.; Drummond, M. L.; Williams, T. G.; Cundari, T. R.; Wilson, A. K. *J. Phys. Chem. A* **2009**, *113*, 8607.
- (7) (a) Gritsenko, O.; Schipper, P. R. T.; Baerends, E. J. *J. Chem. Phys.* **1997**, *107*, 5007. (b) Handy, N. C.; Cohen, A. *J. Mol. Phys.* **2001**, *99*, 403. (c) Cremer, D.; Filatov, M.; Polo, V.; Kraka, E.; Shaik, S. *Int. J. Mol. Sci.* **2002**, *3*, 604.
- (8) Buijse, M. A.; Baerends, E. J. *J. Chem. Phys.* **1990**, *93*, 4129.
- (9) Kahn, L. R.; Baybutt, P.; Truhlar, D. G. *J. Chem. Phys.* **1976**, *65*, 3826.
- (10) Pacios, L. F.; Calzada, P. G. *Int. J. Quantum Chem.* **1988**, *34*, 267.
- (11) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Phys.* **2005**, *123*, 161103.
- (12) (a) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *125*, 194101. (b) Zhao, Y.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 157.
- (13) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215.
- (14) (a) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098. (b) Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
- (15) Chai, J.-D.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615.
- (16) Boese, A. D.; Handy, N. C. *J. Chem. Phys.* **2002**, *116*, 9559.
- (17) (a) Gill, P. M. W. *Mol. Phys.* **1996**, *89*, 433. (b) Adamo, C.; Barone, V. *J. Comput. Chem.* **1998**, *19*, 418.
- (18) Adamo, C.; Barone, V. *J. Chem. Phys.* **1998**, *108*, 664.
- (19) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623.
- (20) Xu, X.; Goddard, W. A., III *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 2673.
- (21) (a) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865. (b) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1997**, *78*, 1396.
- (22) (a) Douglas, M.; Kroll, N. M. *Ann. Phys. (NY)* **1974**, *82*, 89. (b) Hess, B. A. *Phys. Rev. A* **1985**, *32*, 756. (c) Hess, B. A. *Phys. Rev. A* **1986**, *33*, 3742. (d) Jansen, G.; Hess, B. A. *Phys. Rev. A* **1989**, *39*, 6016. (e) Barysz, M.; Sadlej, A. J. *J. Mol. Struct. (Theochem)* **2001**, *573*, 181. (f) de Jong, W. A.; Harrison, R. J.; Dixon, D. A. *J. Chem. Phys.* **2001**, *114*, 48.
- (23) Balabanov, N. B.; Peterson, K. A. *J. Chem. Phys.* **2005**, *123*, 064107.
- (24) Dolg, M.; Wedig, U.; Stoll, H.; Preuss, H. *J. Chem. Phys.* **1987**, *86*, 866.
- (25) Stevens, W. J.; Krauss, M.; Basch, H.; Jasien, P. G. *Can. J. Chem.* **1992**, *70*, 612.
- (26) Hurley, M. M.; Pacios, L. F.; Christiansen, P. A.; Ross, R. B.; Ermler, W. C. *J. Chem. Phys.* **1986**, *84*, 6840.
- (27) (a) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 270. (b) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 284. (c) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 299.
- (28) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297.
- (29) Zheng, J.; Xu, X.; Truhlar, D. G. *Theor. Chem. Acc.* **2011**, *128*, 295.
- (30) (a) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007. (b) Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1993**, *98*, 1358. (c) Wilson, A. K.; Woon, D. E.; Peterson, K. A.; Dunning, T. H., Jr. *J. Chem. Phys.* **1999**, *110*, 7667.
- (31) (a) Hohenberg, P.; Kohn, W. *Phys. Rev. B* **1964**, *136*, 864. (b) Kohn, W.; Sham, L. J. *Phys. Rev. A* **1965**, *140*, 1133. (c) Slater, J. C. *The Self-Consistent Field for Molecular and Solids, Quantum Theory of Molecular and Solids*; McGraw-Hill: New York, 1974; Vol. 4. (d) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200.
- (32) (a) Heyd, J.; Scuseria, G. E. *J. Chem. Phys.* **2004**, *121*, 1187. (b) Heyd, J.; Scuseria, G. E. *J. Chem. Phys.* **2004**, *120*, 7274. (c) Heyd, J.; Peralta, J. E.; Scuseria, G. E.; Martin, R. L. *J. Chem. Phys.* **2005**, *123*, 174101. (d) Heyd, J.; Scuseria, G. E.; Ernzerhof, M. *J. Chem. Phys.* **2006**, *124*, 219906. (e) Izmaylov, A. F.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **2006**, *125*, 104103. (f) Krukau, A. V.; Vydrov, O. A.; Izmaylov, A. F.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 224106. (g) Henderson, T. M.; Izmaylov, A. F.; Scalmani, G.; Scuseria, G. E. *J. Chem. Phys.* **2009**, *131*, 044108.
- (33) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, Revision A.02; Gaussian, Inc.: Wallingford, CT, 2009.
- (34) Gutsev, G. L.; Bauschlicher, C. W., Jr. *J. Phys. Chem. A* **2003**, *107*, 4755 and references therein.
- (35) Andersson, N.; Balfour, W. J.; Bernath, P. F.; Lindgren, B.; Ram, R. S. *J. Chem. Phys.* **2003**, *118*, 3543.
- (36) Bauschlicher, C. W., Jr.; Maitre, P. *Theor. Chim. Acta* **1995**, *90*, 189.
- (37) Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1977**, *66*, 3045.
- (b) Bauernschmitt, R.; Ahlrichs, R. *J. Chem. Phys.* **1996**, *104*, 9047.
- (c) Schlegel, H. B.; McDouall, J. J. *Computational Advances in Organic Chemistry*; Ögretir, C., Csizmadia, I. G., Eds.; Kluwer Academic: The Netherlands, 1991; p 167.
- (38) Martin, J. M. L.; Sundermann, A. *J. Chem. Phys.* **2001**, *114*, 3408.
- (39) Metz, B.; Stoll, H.; Dolg, M. *J. Chem. Phys.* **2000**, *113*, 2563.
- (40) Sansonetti, J.; Martin, W. Young, S. *Handbook of Basic Atomic Spectroscopic Data*; Version 1.00, National Institute of Standards and Technology: Gaithersburg, MD, 2003. Available at <http://physics.nist.gov/Handbook> (accessed Aug 10, 2011).

Fully Relativistic Calculations of Faraday and Nuclear Spin-Induced Optical Rotation in Xenon

Suvi Ikäläinen,^{*,†} Perttu Lantto,[‡] and Juha Vaara[‡][†]Laboratory of Physical Chemistry, Department of Chemistry, P.O. Box 55 (A.I. Virtasen aukio 1), FIN-00014 University of Helsinki, Finland[‡]NMR Research Group, Department of Physics, P.O. Box 3000, FIN-90014 University of Oulu, Finland Supporting Information

ABSTRACT: Nuclear spin-induced optical rotation (NSOR) arising from the Faraday effect may constitute an advantageous novel method for the detection of nuclear magnetization. We present first-principles nonrelativistic and relativistic, two- and four-component, basis-set limit calculations of this phenomenon for xenon. It is observed that only by utilization of relativistic methods may one qualitatively reproduce experimental liquid-state NSOR data. Relativistic effects lower the results by 50% as compared to nonrelativistic values. Indeed, relativistic Hartree–Fock calculations at the four-component or exact two-component (X2C) level account for the discrepancy between experimental results and earlier nonrelativistic theory. The nuclear magnetic shielding constant of traditional nuclear magnetic resonance as well as the Verdet constant parametrizing optical rotation due to an external magnetic field were also calculated. A comparison between results obtained using Hartree–Fock and density-functional theory methods at relativistic and nonrelativistic levels, as well as coupled cluster methods at the nonrelativistic level, was carried out. Completeness-optimized basis sets were employed throughout, for the first time in fully relativistic calculations. Full relativity decreases the Verdet constant by 4%. X2C theory decreases the absolute value of NSOR by 10–20% as compared to the four-component data, while for Verdet constants, the results are only slightly smaller than the fully relativistic values. For both properties, two-component calculations decrease the computational time by roughly 90%. Density-functional methods yield substantially larger values of NSOR than the Hartree–Fock theory or experiments. Intermolecular interactions are found to decrease NSOR and, hence, compensate for the electron correlation effect.

1. INTRODUCTION

Optical phenomena have been discussed in many recent studies^{1–12} as novel methods for the detection of nuclear magnetic resonance (NMR). Optical effects may be more readily observed and carry the possibility of enhanced spatial resolution as compared to traditional radio frequency detection. These phenomena are based on the Faraday and inverse Faraday effects. In the Faraday effect, a magnetic field causes the plane of polarization of a linearly polarized (LPL) light beam, directed along the field, to rotate.¹³ The field due to spin-polarized nuclei in an NMR sample accordingly also causes rotation in the plane of polarization of the incident LPL, in what is called the nuclear spin-induced optical rotation (NSOR).⁹ Arising from the opposite phenomenon of the inverse Faraday effect, the laser-induced NMR shift has been investigated theoretically in refs 1–8, of which refs 6–8 report first-principles calculations of the magnitude of the phenomenon. It has been seen that, at least far away from the immediate vicinity of optical resonances, the effect is far too small to be measured. The NSOR has, however, been observed experimentally for protons in liquid water and liquid ¹²⁹Xe in ref 9 and ¹⁹F in fluorocarbons in ref 12. First-principles nonrelativistic (NR) theoretical evaluation of NSOR has been carried out for ethanol, nitromethane, water, and urea in ref 11, where excellent agreement with experimental results was achieved for liquid water. Chemical distinction between different molecules and inequivalent nuclei in the same molecule has been predicted¹¹

and observed,¹² which implies that NSOR could provide a viable and potentially more informative analogue to the NMR chemical shift of traditional NMR detection. For ethanol, immense amplification of the effect was observed at laser frequencies close to optical resonances.¹¹ Enhanced chemical distinction between optically excited chromophores in the light-sensitive retinal model PSB-11 was also demonstrated.¹¹ A study regarding the intermolecular interaction effects on the NSOR due to ¹H and ¹⁷O nuclei in H₂O(l) has also been conducted recently,¹⁴ demonstrating a close cancellation of the solvation and local optical field effects for ¹HSOR.

The laser-induced shift and the NSOR are calculated through a similar third-order perturbational expression, essentially that of molecular antisymmetric polarizability,^{1,2} and are easily interconverted.^{9,11} In ref 9, the experimental results for liquid ¹²⁹Xe were compared to NR theoretical values of the laser-induced shift obtained in ref 7. A qualitative agreement was observed, but it was subsequently realized that a factor of 2 had been neglected in the theory of ref 7, destroying the compatibility of the results. This implies that relativistic effects may be relevant for NSOR in ¹²⁹Xe.

The importance of relativistic effects for accurate calculations of NSOR for heavy nuclei arises from the quantum mechanical

Received: September 12, 2011

Published: November 08, 2011

hyperfine operator that is involved in the antisymmetric polarizability, accountable for NSOR. In addition to the demands placed on the description of the electronic structure of the inner shells due to the hyperfine operator, the calculation of both NSOR and the Verdet constant (parametrizing the conventional Faraday rotation due to an external magnetic field) requires accurate electronic structure also at the outskirts of the electronic cloud, as this region is readily distorted by the external electric field, with significant contributions to the polarizability. Basis sets of high quality must therefore be used, which leads us to utilization of the completeness optimization¹⁵ to generate basis sets designed specifically for the basis-set limit calculation of NSOR in ¹²⁹Xe.

Here, we aim to demonstrate the importance and magnitude of relativistic effects on ¹²⁹Xe NSOR in gaseous and liquid xenon through fully relativistic (strictly true only for the one-particle part of the molecular Hamiltonian¹⁶) four-component Dirac–Fock (DHF) and Dirac–DFT (DDFT) calculations. ¹²⁹XeNSOR has been calculated at standard visible or near-infrared laser frequencies. We have also evaluated the Verdet constant for xenon as well as, to further evaluate the basis set used, the nuclear shielding constant of traditional NMR. The DHF results are compared to relativistic, sc. exact two-component (X2C)¹⁷ data as well as NR HF results. Relativistic and NR calculations were carried out at the HF and density functional theory (DFT) levels, and both the collinear and the noncollinear spin density definitions were used in the relativistic code for the latter. NR coupled cluster singles and doubles (CCSD) values were computed to calibrate the performance of the DFT functionals. Novel and compact completeness-optimized basis sets, which have been shown to produce results close to the basis-set limit for magnetic properties,^{8,11,15,18} were employed here in a first application to fully relativistic calculations. Basis-set convergence for the nuclear shielding constant in atomic ¹²⁹Xe is also of present interest, as it earlier turned out to be demanding by a Gaussian basis-set expression^{19–21} to approach the numerical limiting value.²²

2. THEORY

2.1. Nuclear Spin-Induced Optical Rotation. The magnetic optical rotation angle Φ per unit of sample length l can be written as^{23–25}

$$\frac{\Phi}{l} = \frac{1}{2} \omega \mathcal{N} \mu_0 c \text{Im} \langle \alpha'_{XY} \rangle \quad (1)$$

where ω is the frequency of the laser beam propagating in the laboratory Z direction, \mathcal{N} is the number density, c denotes the speed of light *in vacuo*, and $\langle \alpha' \rangle$ is the ensemble-averaged, complex antisymmetric polarizability. For an external magnetic field \mathbf{B}_0 and nuclear spin component $I_{K,Z}$ along the beam^{1,10}

$$\alpha'_{XY} = \alpha'_{XY,Z} B_0 + \alpha'_{XY,Z} I_{K,Z} + \mathcal{O}(B_0^3, I_K^3) \quad (2)$$

Molecular tumbling in gaseous or liquid samples leads to the isotropic molecular average

$$\begin{aligned} \langle \alpha'_{XY,Z} \rangle &= \frac{1}{6} \sum_{\varepsilon\tau\nu} \varepsilon_{\varepsilon\tau\nu} \alpha'_{\varepsilon\tau,\nu} \\ &= \frac{1}{3} (\alpha'_{xy,z} + \alpha'_{yz,x} + \alpha'_{zx,y}) \end{aligned} \quad (3)$$

where $\varepsilon_{\varepsilon\tau\nu}$ is the Levi–Civita symbol and (x,y,z) are coordinates in the molecule-fixed Cartesian frame. For an atom, the three

components are equal and $\langle \alpha'_{XY,Z} \rangle = \alpha'_{xy,z} = \alpha'^{(B_0/I_K)}_{\varepsilon\tau,\nu}$ may be calculated through a third-order perturbation theoretical equation^{1,2} that, in the notation of quadratic response theory,^{6,26} is written as

$$\alpha'^{(B_0/I_K)}_{\varepsilon\tau,\nu} = - \langle \langle \mu_\varepsilon; \mu_\tau, h_\nu^{Z/\text{hf}} \rangle \rangle_{\omega,0} \quad (4)$$

with μ_ε and μ_τ arising from the components of the dipole moment. In eq 4, the expression of conventional electric dipole polarizability, $\alpha(\omega) = - \langle \langle \mu; \mu \rangle \rangle_{\omega}$, is modified by a third, static magnetic operator h . For optical rotation caused by an external field \mathbf{B}_0 , this operator is the Zeeman interaction, whereas for NSOR, h is the hyperfine interaction. The relativistic perturbation operators are defined through

$$H_Z = \sum_\nu h_\nu^Z B_{0,\nu}; \quad H_{\text{hf}} = \sum_K \sum_\nu h_{K,\nu}^{\text{hf}} I_{K,\nu} \quad (5)$$

for the Zeeman and hyperfine interactions, respectively, where

$$h_\nu^Z = - \frac{ce}{2} \sum_{i=1}^{N_d} (\boldsymbol{\alpha} \times \mathbf{r}_{iO})_\nu \quad (6)$$

and

$$h_{K,\nu}^{\text{hf}} = - \frac{ce\mu_0 \hbar}{4\pi} \sum_{i=1}^{N_d} \frac{\gamma_K (\boldsymbol{\alpha} \times \mathbf{r}_{iK})_\nu}{r_{iK}^3} \quad (7)$$

Here, $\boldsymbol{\alpha}$ is the Dirac 4×4 matrix operator, N_d is the number of electrons, γ_K is the gyromagnetic ratio of nucleus K , and \mathbf{r}_{iK} the vector from the nucleus K . In one-component NR theory, the familiar orbital Zeeman (OZ) and paramagnetic nuclear spin–electron orbit (PSO) operators¹¹ replace the operators of eqs 6 and 7, respectively.

In the case of optical rotation caused by an external field \mathbf{B}_0 , $\Phi = VB_0 l$, where the Verdet constant V is

$$V = - \frac{1}{2} \omega \mathcal{N} \mu_0 c e^2 \frac{1}{6} \sum_{\varepsilon\tau\nu} \varepsilon_{\varepsilon\tau\nu} \text{Im} \langle \langle r_\varepsilon; r_\tau, h_\nu^Z \rangle \rangle_{\omega,0} \quad (8)$$

For optical rotation arising from nuclear spins, for unit concentration $[\] = \mathcal{N}/N_A$ of the polarized nuclei K ,

$$\frac{\Phi_{\text{NSOR}}}{[l]} = - \frac{1}{2} \omega N_A \mu_0 c e^2 \langle I_{K,Z} \rangle \frac{1}{6} \sum_{\varepsilon\tau\nu} \varepsilon_{\varepsilon\tau\nu} \text{Im} \langle \langle r_\varepsilon; r_\tau, h_{K,\nu}^{\text{hf}} \rangle \rangle_{\omega,0} \quad (9)$$

where $\langle I_{K,Z} \rangle$ is the average spin polarization.

The NSOR and the laser-induced shift, Δ , can be related through the equation

$$\frac{\Phi_{\text{NSOR}}}{[l]} = - h \omega N_A \langle I_{K,Z} \rangle \frac{\Delta}{I_0} \quad (10)$$

where I_0 is the intensity of the incident, circularly polarized beam in the inverse Faraday effect.

2.2. Completeness-Optimized Basis Sets. The concept of completeness optimization was first introduced by Manninen and Vaara in ref 15 as a novel method of generating high-quality Gaussian basis sets with relatively few functions. In the completeness-optimization scheme, energetic criteria for basis-set generation are discarded allowing, in principle, creation of universal (element-independent) basis sets that are systematically and economically extended for basis-set limit calculations of a specific property. Completeness profiles presented by Chong²⁷ are employed. The completeness profile is defined as

$$Y(\zeta) = \sum_m \langle g(\zeta) | \chi_m \rangle^2 \quad (11)$$

where $\{\chi\}$ is a set of orthonormalized basis functions for a given angular momentum l and $g(\zeta)$ is an arbitrary “test” Gaussian l -type orbital (GTO) with the exponent ζ . $g(\zeta)$ is used to analyze the completeness of $\{\chi\}$, and for a complete set, the value of $Y(\zeta)$ is equal to 1 for all ζ . $Y(\zeta)$ can be portrayed on a $[\log(\zeta), Y(\zeta)]$ plot, in which case the profile of a basis set that is complete for a certain range of ζ will create a plateau-like figure, an example of which may be seen in the Supporting Information, where the completeness profile for the basis set used presently is displayed. Completeness-optimized basis sets are generated using the Kruunuhaka code,²⁸ in which one may specify the desired exponent range $[\zeta_{\min}, \zeta_{\max}]$ and the number of GTOs. The code will then generate a primitive basis set using these criteria, for which the measure of the deviation from completeness¹⁵

$$\tau = \int_{\zeta_{\min}}^{\zeta_{\max}} [1 - Y(\zeta)] d\zeta \quad (12)$$

will be as small as possible, resulting in a compact basis set that will typically produce more accurate results for magnetic properties than traditional energy-optimized basis sets of the same size.^{8,15,18}

3. CALCULATIONS

Completeness optimization was used to generate the large-component (LC) basis sets. The small-component (SC) basis sets used in the relativistic calculations were obtained via restricted kinetic balance (RKB) or unrestricted kinetic balance (URKB).¹⁶ The basis sets were obtained by first generating a set that, for the different l values, spans the same exponent ranges as in the cv4z basis of Dyall,²⁹ maintaining the deviation τ from completeness [eq 12] under 0.001. The exponent ranges were then systematically extended by tight and diffuse exponents, for each l value separately, using a sufficient number of Gaussian functions to span the desired area to high accuracy. Trial DHF calculations of ¹²⁹Xe nuclear shielding and ¹²⁹XeSOR were conducted, and an exponent range that no longer significantly changed the results was chosen as a reference. The number of exponents was then reduced for each of the l values separately, and the basis set giving results for the nuclear shielding constant and NSOR deviating by under 0.5 ppm and 1%, respectively, from the reference values, was chosen as the “co” basis set with the final composition (35s32p24d3f). The exponents and completeness profile of this basis set are available in the Supporting Information.

All relativistic calculations of Φ_{NSOR} , V , and the nuclear shielding constant σ were conducted with the Dirac³⁰ program, while all NR calculations were carried out with the Dalton program.³¹ Verdet constants are reported for both gaseous and liquid Xe. The gas number density \mathcal{N} required for the conversion of the calculated single-atom response functions to real gas situation was obtained through the van der Waals equation for real xenon gas,³² which yields $\mathcal{N} = 2.58 \times 10^{25} \text{ m}^{-3}$ at STP. Calculations were done at standard UV/NIR laser frequencies using HF and different DFT methods as well as NR CCSD. The DFT functionals BLYP,^{33,34} B3LYP,^{34–36} and BHandHLYP,^{34,37} with increasing amounts of the exact HF exchange (0%, 20%, and 50%, respectively) were employed. The four-component relativistic calculations of nuclear shielding were carried out using full linear response theory, i.e., without replacing the ep branch of the response function with the corresponding NR diamagnetic operator, as is sometimes done in four-component theory.^{19,38} In addition to four-component calculations, two-component (X2C)¹⁷ results were also computed for Φ_{NSOR} and V . X2C values are not reported

for the nuclear shielding constant, as the diamagnetic part of the shielding is not yet properly accounted for in the X2C scheme of the Dirac program. Quadratic response functions were used for HF, DFT, and CCSD, the implementations of which are covered in refs 39, 40, and 41, respectively, for the Dalton program, and refs 42 and 43 for the Dirac code.

The basis-set convergence of Dyall’s vxz and cvxz^{29,44,45} basis set families was investigated for Φ_{NSOR} , V , and σ with the HF method. The effects of using either RKB or URKB for the generation of the SC basis sets was also investigated. The completeness-optimized basis set was then used for calculations of all of the discussed properties with the DHF, X2C (omitting σ), and NR HF methods to determine relativistic effects. The co basis set was also used at the different DFT levels (at the four-component and NR levels), for which relativistic calculations using both the non-collinear⁴⁶ and collinear definitions of the spin density were also performed. All calculations involving the co basis were carried out with URKB.

The effects of numerical approximations as well as the inclusion and exclusion of two-electron integral classes were tested comprehensively, but results are not given here, as the deviation between the results was minimal. The convergence thresholds for the wave function and the response functions were dropped by a factor of 10 from the Dirac defaults in separate calculations. A convergence threshold of 1.2×10^{-5} was used for the wave function optimization, while 1.0×10^{-7} was used for both linear and quadratic response functions. A calculation using the full set on two-electron integrals as well as calculations excluding the small–small (SS) integrals from the response, excluding both the SS and large–small (LS) integrals from the response, as well as a calculation excluding the SS integrals from both the response and wave function optimization, were also conducted. The reported results are obtained via excluding SS integrals in the response as well as SCF. The same approximation was found to be entirely adequate for Verdet constants in ref 47.

4. RESULTS AND DISCUSSION

4.1. Basis-Set Convergence with Standard Sets. The basis set convergence of the NSOR angle and the Verdet constant at 514.5 nm, as well as the nuclear shielding constant as a function of the number of basis functions (LC functions for relativistic results) of the Dyall basis set families is shown in Figure 1. The DHF URKB results given by the co basis set are also displayed. Tables S2–S4 in the Supporting Information list $\Phi_{\text{NSOR}}/([\text{I}])$ and V at the different laser wavelengths, along with σ , using the Dyall basis set families and same levels of theory as in Figure 1. RKB results are omitted from the tables for NSOR and V , as they are identical to the URKB values to the displayed accuracy. Table 1 gives the number of basis functions (LC and SC) for the Dyall and co basis sets.

As in previous studies,^{9,11} it is seen that the magnitude of $\Phi_{\text{NSOR}}/([\text{I}])$ decreases as the laser wavelength increases. The difference between RKB and URKB data is practically negligible for NSOR, and dyall.vxz and dyall.cvzx results are very close to each other, although the addition of tight functions in the cvzx series increases NSOR slightly. The inclusion of relativity decreases absolute values of $\Phi_{\text{NSOR}}/([\text{I}])$ by roughly 15–30%. It is seen that the results with the two Dyall basis set series are not converged as a function of the basis set size. Furthermore, they are monotonically increasing, away from the basis-set limit value obtained with the presently developed co basis set. Indeed,

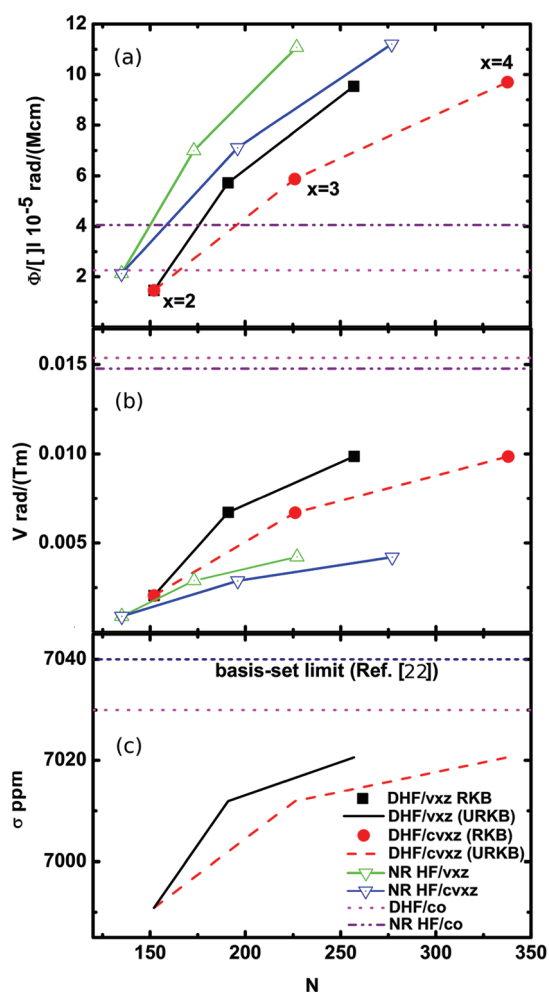


Figure 1. (a) ^{129}Xe nuclear spin optical rotation angle $\Phi_{\text{NSOR}}/(|I|)$ [in 10^{-5} rad/(M cm)], (b) the Verdet constant V [in rad/(T m)] (both properties at 514.5 nm), as well as (c) the nuclear shielding constant σ [in ppm] for gaseous Xe as a function of the number of basis functions N using the Dyllal vxz and cvzx basis sets and the nonrelativistic (NR) as well as relativistic (DHF) Hartree–Fock method. Both the restricted and unrestricted kinetic balance (RKB and URKB, respectively) were used for the latter. The results obtained with the co basis set (with $N = 305$) are shown as horizontal lines. Due to the large offset in results, the NR and DHF/RKB data for σ are not shown.

Table 1. Number of Large-Component (LC) and Small-Component (SC) Basis Functions in the Dyllal and co Basis Sets

basis set	LC	SC		total	
		RKB	URKB	RKB	URKB
dyall.v2z	152	177	353	329	505
dyall.v3z	191	224	447	415	638
dyall.v4z	257	298	595	555	852
dyall.cv3z	226	260	520	486	746
dyall.cv4z	338	382	763	720	1101
co	305		704		1009

augmentation of the dyall.v4z basis with diffuse d -type functions (obtained by successively dividing the most diffuse exponent by 3)

Table 2. ^{129}Xe Nuclear Spin Optical Rotation $\Phi_{\text{NSOR}}/(|I|)$ [in 10^{-5} rad/(M cm)] for Different Laser Wavelengths at the Hartree–Fock Level Using the Completeness-Optimized Basis Set co and the Fully Relativistic Four-Component (DHF), Exact Two-Component (X2C), and Nonrelativistic (NR) Methods (Experimental Results Are Also Given)

λ (nm)	ω (au)	NR	X2C	DHF	exptl. ^a
488.8	0.0932147	−4.65	−2.30	−2.63	
514.5	0.0885585	−4.05	−1.97	−2.25	
532.0	0.0856454	−3.71	−1.78	−2.04	1.5 ± 0.3
589.0	0.0773571	−2.87	−1.32	−1.53	
694.3	0.0656249	−1.92	−0.84	−0.99	
770.0	0.0591732	−1.51	−0.64	−0.76	0.6 ± 0.1
1064.0	0.0428227	−0.74	−0.29	−0.35	0.4 ± 0.2
1319.0	0.0345439	−0.47	−0.18	−0.22	

^a Liquid-state experimental results from ref 9.

Table 3. Verdet Constant V [in 10^{-3} rad/(T m)] for Gaseous Xe at Different Laser Wavelengths Using the Completeness-Optimized Basis Set co and the Fully Relativistic Four-Component (DHF), Exact Two-Component (X2C), and Nonrelativistic (NR) Hartree–Fock Methods (Previous Relativistic Computational and Experimental Results Are Also Given)

λ (nm)	ω (au)	NR	X2C	DHF	ref 47 ^a	exptl. ^b
488.8	0.0932147	16.53	17.31	17.23		
514.5	0.0885585	14.77	15.44	15.37	15.59	
532.0	0.0856454	13.73	14.35	14.28		
589.0	0.0773571	11.02	11.49	11.44		12.30
694.3	0.0656249	7.77	8.09	8.06	8.15	
770.0	0.0591732	6.26	6.51	6.48		
1064.0	0.0428227	3.21	3.34	3.32	3.37	
1319.0	0.0345439	2.08	2.15	2.15		

^a DHF results using the well tempered-basis set by Huzinaga augmented with diffuse functions. Values in the table have been converted to real gas number density. Original values obtained in ref 47 are 55.16, 28.85, and 11.91 $\mu\text{min}/(\text{G cm})$ at $\omega = 0.088599$, 0.065600, and 0.042823 au, respectively, for an ideal gas. ^b Experimental result from ref 48.

alters the results (not shown) dramatically toward the co values. As before,^{8,11} the antisymmetric polarizability responsible for NSOR is very challenging even for high-quality basis sets designed for studying standard chemical problems.

Similarly to NSOR, the magnitude of V increases with frequency, and RKB and URKB data are in practice equivalent. Additional functions in the dyall.cvzx sets do not particularly affect the results, which is expected, as no hyperfine operators are involved. For the same reason, the NR results are only 3–5% lower than the relativistic results, as the response of the valence-only operators of V is less sensitive to changes in the description of the atomic core region than in the case of NSOR. As in earlier work,⁴⁷ the Verdet constants are underestimated by basis sets that lack sufficiently diffuse functions.

From Table S4 (Supporting Information), it is seen that for σ , the use of URKB clearly improves the results and is therefore necessary to approach the basis-set limit with sets of the size of the Dyllal basis set families. In the URKB case, the more complete small-component basis set is important for the part of

corresponding to negative-energy excited states. In calculations with URKB, the magnitude of σ increases by 400–500 ppm as compared to RKB results, and the series starts to show signs of convergence. The NR data are approximately 1400 ppm lower than the relativistic URKB data. In contrast to the case of NSOR, additional diffuse d functions on top of the Dyall v4z basis do not affect σ . The best results with the Dyall basis sets are approximately 10 ppm lower than the co result and 20 ppm below the basis-set limit of 7040 ppm.²²

4.2. Four-Component, Two-Component, and NR Calculations. We turn to nearly basis-set limit calculations using our present co basis set. Tables 2 and 3 display NSOR and the Verdet constant as functions of the laser wavelength at the fully relativistic four-component DHF, two-component X2C, and NR

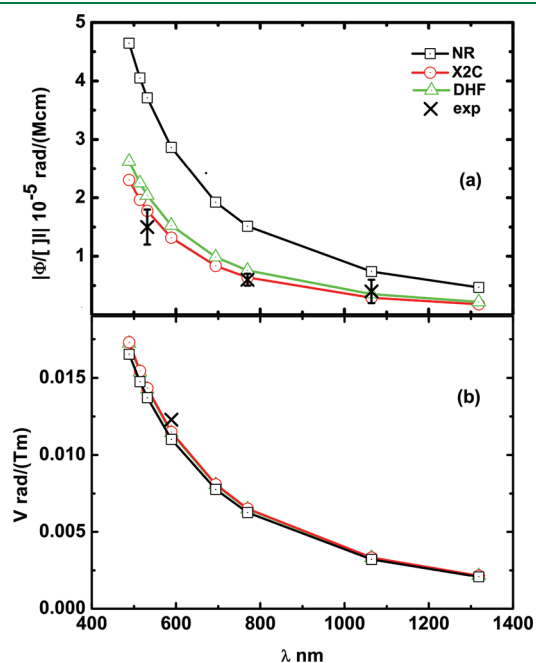


Figure 2. (a) ^{129}Xe nuclear spin optical rotation $|\Phi_{\text{NSOR}}| / (l)$ [in $10^{-5} \text{ rad}/(\text{M cm})$] and (b) Verdet constant V $\text{rad}/(\text{T m})$ for gaseous Xe at different laser wavelengths using the completeness-optimized basis set co and the fully relativistic four-component (DHF), exact two-component (X2C), and nonrelativistic (NR) Hartree-Fock methods. The experimental data from refs 9 and 48 are also shown.

HF levels, with the co basis set. The corresponding nuclear shielding data are included in Table S4 (Supporting Information). Experimental results are also reported for NSOR and V . Figure 2 illustrates $|\Phi_{\text{NSOR}}| / (l)$ and V , the latter for gaseous Xe. Going from the NR calculations to full relativity decreases the absolute value of $^{129}\text{XeSOR}$ by roughly 40–50%. DHF results are in reasonably good agreement with the experimental values taking the error limits of the latter into account. Table 2 shows that the good correspondence between the NR calculations of ref 7 with experimental results discussed in ref 9 was indeed due to the absence of a factor of 2 in the analysis of ref 7. The present inclusion of relativity brings the results back close to the experimental values. The X2C method further reduces NSOR as compared to DHF, producing (fortuitously) a still slightly improved agreement with experimental results.

For the Verdet constant, the NR results are below the DHF data by $\sim 4\%$, whereas the X2C calculations yield slightly larger values still. The DHF and X2C results are close to the experimental⁴⁸ value at $\lambda = 589.0 \text{ nm}$ as well as the previous computational DHF results⁴⁷ obtained with an augmented well-tempered Huzinaga basis set.

4.3. Correlation Effects with the DFT Method. Table 4 and Tables S5 and S6 in the Supporting Information give NSOR, V , and σ at HF and various DFT levels of theory, for both noncollinear and collinear treatment of the spin density in the case of DDFT. NR CCSD results are also displayed. Figure 3 illustrates NSOR and the Verdet constant as a function of wavelength with the HF and DFT methods. For all properties, it is seen that the collinear and noncollinear approaches give nearly identical results, and the numerical values indicate that the small difference further diminishes with increasing exact exchange admixture in the DFT functional in the series from BLYP via B3LYP to BHandHLYP. The increase in the amount of exact exchange leads to an overall decrease in $|\Phi_{\text{NSOR}}|$ toward the HF values. The DFT results remain significantly larger than the HF results, by $\sim 50\text{--}100\%$ and $70\text{--}300\%$ at the NR and four-component levels, respectively. A smaller variation between the HF and DFT results is observed for V , for which DFT leads to increases of $\sim 10\text{--}40\%$ and $\sim 15\text{--}50\%$ at NR and relativistic levels. The DFT data are typically also in a much greater disagreement with experimental results than the HF results, with the exception of V at the BHandHLYP level, which is already rather close to the experimental result. Compared to NR HF, electron correlation at the NR CCSD level increases $^{129}\text{XeSOR}$ by $\sim 65\text{--}75\%$, whereas a smaller relative

Table 4. ^{129}Xe Nuclear spin-induced optical rotation $\Phi_{\text{NSOR}} / (l)$ [in $10^{-5} \text{ rad}/(\text{M cm})$] at different laser wavelengths using the completeness-optimized basis set co with nonrelativistic (NR) as well as relativistic (Rel.) HF and different DFT methods. Both the non-collinear (NC) and collinear (C) spin density approaches were used for relativistic DFT. The nonrelativistic coupled-cluster singles and doubles (CCSD) results are also given.

γ (nm)	ω (a.u.)	DFT/BLYP			DFT/B3LYP			DFT/BHandHLYP			HF		NR CCSD
		NR	Rel. C	Rel. NC	NR	Rel. C	Rel. NC	NR	Rel. C	Rel. NC	NR	Rel.	
488.8	0.0932147	-14.60	-11.32	-11.24	-10.70	-7.73	-7.68	-7.18	-4.54	-4.54	-4.65	-2.63	-7.65
514.5	0.0885585	-12.77	-9.86	-9.79	-9.36	-6.73	-6.68	-6.28	-3.93	-3.93	-4.05	-2.25	-6.71
532.0	0.0856454	-11.71	-9.03	-8.96	-8.59	-6.16	-6.12	-5.76	-3.59	-3.59	-3.71	-2.04	-6.17
589.0	0.0773571	-9.09	-6.96	-6.91	-6.67	-4.74	-4.71	-4.47	-2.74	-2.74	-2.87	-1.53	-4.82
694.3	0.06562496	-6.15	-4.67	-4.64	-4.52	-3.17	-3.15	-3.02	-1.82	-1.82	-1.92	-0.99	-3.28
770.0	0.0591732	-4.85	-3.67	-3.64	-3.57	-2.49	-2.48	-2.39	-1.42	-1.42	-1.51	-0.76	-2.60
1064.0	0.0428227	-2.39	-1.79	-1.78	-1.76	-1.21	-1.21	-1.17	-0.68	-0.68	-0.74	-0.35	-1.29
1319.0	0.0345439	-1.52	-1.14	-1.13	-1.12	-0.77	-0.76	-0.75	-0.43	-0.43	-0.47	-0.22	-0.82

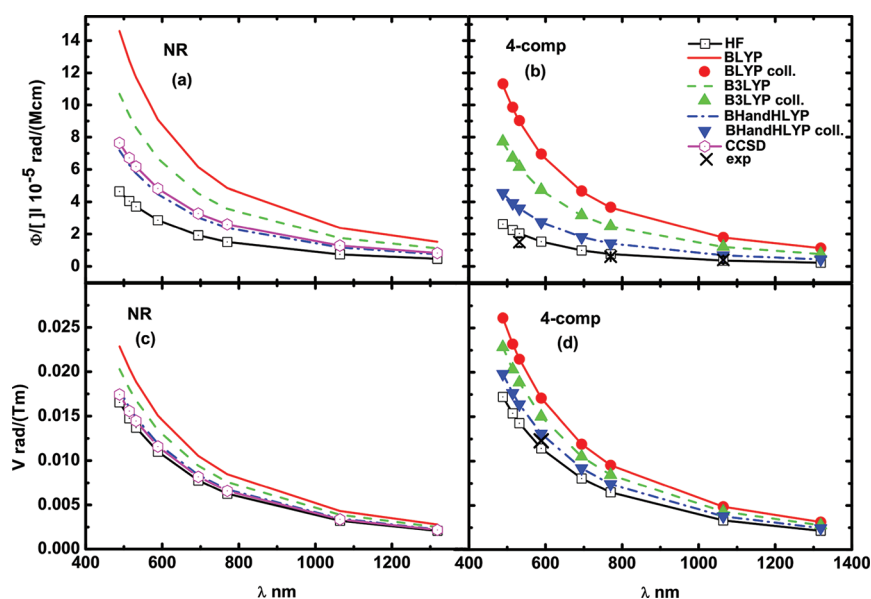


Figure 3. (a,b) ^{129}Xe nuclear spin-induced optical rotation angle [in 10^{-5} rad/(M cm)] and (c,d) the Verdet constant [in rad/(T m)] for gaseous Xe at different laser wavelengths using the completeness-optimized basis set co and the Hartree–Fock (HF) method as well as density functional theory with BLYP, B3LYP, and BHandHLYP functionals. Nonrelativistic results are displayed on the left (a and c), while the relativistic results are given on the right (b and d).

increase of 5% is observed for V . For both properties at the NR level, BHandHLYP is closest to CCSD results. From the similar behavior of NR DFT and DDFT, and from the increase in NR $|\Phi_{\text{NSOR}}|$ in Figure 3 upon introduction of electron correlation, it can be conjectured that electron correlation would also increase the absolute values of $^{129}\text{XeSOR}$ at the relativistic level, taking them further away from the experimental values. We presently lack the computational tools for correlated *ab initio* magnetic properties at the fully relativistic level. For σ , the HF and DFT values are very similar to each other.

4.4. Comparison with Experimental Results. It was seen in Figure 2 that inclusion of relativity brings the HF results for NSOR close to experimental values. However, it is also evident from Table 4 that electron correlation, also at the relativistic level (as estimated using DDFT), again renders these results further from the experimental ones. It seems as though our results for $^{129}\text{XeSOR}$ based on isolated-atom calculations remain higher than the experimental values. The latter were obtained in liquid Xe, and the concentration of the atoms provides the connection of optical rotation, a bulk property, to the calculated antisymmetric polarizability of a single atom. Atomic and molecular properties change, however, when they are introduced to a medium. It was found in ref 14 that, for ^1H in water, Φ_{NSOR} for an interacting molecule in the liquid phase is 14% smaller than NSOR for a static molecule *in vacuo*. For the oxygen nucleus, the NSOR is 29% smaller for interacting molecules. It is thus likely that $^{129}\text{XeSOR}$ for liquid-phase Xe would also be lower than our *in vacuo* results. This was tested by performing a DHF calculation on a ^{129}Xe dimer at its equilibrium geometry [$r_{\text{Xe}-\text{Xe}} = 4.3627$ Å (ref 49)] with the co basis set, the results of which are reported in Table S7 of the Supporting Information. Although the co basis is not entirely converged for the interaction effect, it can be seen that the values of NSOR are indeed lowered by ~ 35 – 45% . Hence, interatomic interactions are important for $^{129}\text{XeSOR}$, and their proper inclusion is likely to significantly improve the agreement with experimental optical rotation in a liquid medium.

5. CONCLUSIONS

Fully relativistic calculations of the nuclear spin-induced optical rotation at standard vis–near-IR laser frequencies were conducted for ^{129}Xe , along with computations of the Verdet constant and nuclear shielding. Completeness optimization, a novel method for generation of basis sets that have been proven to be successful in calculations of magnetic properties, was used for the first time in fully relativistic calculations. The presently generated co basis set was compared to the Dyll basis set families, for which calculations were performed using both restricted and unrestricted kinetic balance. The significance of relativity was evaluated with calculations at fully relativistic four-component Dirac Hartree–Fock, exact two-component HF, and nonrelativistic HF levels of theory. Various DFT functionals were also utilized at NR and relativistic levels, for which both the collinear and noncollinear spin density approaches were examined. At the NR level, the *ab initio* CCSD method was used as a benchmark for electron correlation effects.

It was observed that the Dyll basis sets appear to converge to an erroneous basis-set limit for the present, very demanding property of NSOR. RKB and URKB give very similar results for both NSOR and Verdet constants. DHF and X2C results are relatively similar for Φ_{NSOR} and V . The relativistic and nonrelativistic HF Verdet constants are close to each other, with relativity adding a few percent to the results, similarly to earlier observations.⁴⁷ The results are close to experimental and previous DHF data. For $^{129}\text{XeSOR}$, full relativity lowers the NR results by 40–50%, while the X2C results remain still somewhat lower than DHF values. The inclusion of relativity is mandatory in order to reach a qualitative agreement with recent $^{129}\text{XeSOR}$ experiments on liquid xenon. Earlier NR calculations by one of the present authors were fortuitously successful due to a missing numerical factor in the analysis.

All of the investigated DFT levels give larger values than HF of $^{129}\text{XeSOR}$ and the Verdet constants in both relativistic and NR

calculations. Electron correlation effects estimated via NR CCSD calculations increase $^{129}\text{XeSOR}$ and V by ca. 70% and 5%, respectively. Among DFT results, BHandHLYP values are closest to experimental ones and, at the NR level, CCSD values, as noted before.⁸ It can be concluded that while the uncorrelated DHF values for NSOR are closer to experimental results than DDFT data, the inclusion of electron correlation does lead to overestimation. These calculations were made for a noninteracting Xe atom, which led us to approximate the intermolecular interaction effects by performing a calculation for a Xe dimer. The results indicate that calculations involving interacting molecules would, in turn, decrease NSOR as compared to isolated molecules, bringing the values back closer to experimental ones. Hence, relativistic, electron correlation, and intermolecular interaction effects are all important for heavy-atom NSOR.

■ ASSOCIATED CONTENT

S Supporting Information. Exponents and the completeness profile of the co basis set; tables of NSOR and Verdet constants (for gaseous ^{129}Xe) at different wavelengths, as well as nuclear shielding using the Dyall basis set families (nonrelativistic and four-component Hartree–Fock methods); tables of NSOR, Verdet constants (for gaseous ^{129}Xe) at different wavelengths, and nuclear shielding using the co basis set at the Hartree–Fock and different DFT levels using nonrelativistic and four-component theory, for which noncollinear and collinear definitions of spin density were used for all of the DFT functionals; table of NSOR for the interacting Xe dimer and noninteracting atom at different wavelengths at the co/DHF level. This material is available free of charge via the Internet at <http://pubs.acs.org/>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: suvi.ikalainen@helsinki.fi.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENT

We thank Prof. Michael V. Romalis (Princeton) for useful discussions. The authors belong to the Finnish CoE in Computational Molecular Science. Support was received from the Graduate School of Computational Chemistry and Molecular Spectroscopy (S.I.), the Alfred Kordelin Fund (S.I.), Academy of Finland (P.L., J.V.), and U. Oulu (J.V.). Computational resources due to CSC (Espoo, Finland) were used. P.L. is an academy research fellow of the Academy of Finland.

■ REFERENCES

- (1) Buckingham, A. D.; Parlett, L. C. *Science* **1994**, *264*, 1748.
- (2) Buckingham, A. D.; Parlett, L. C. *Mol. Phys.* **1997**, *91*, 805.
- (3) Warren, W. S.; Mayr, S.; Goswamy, D.; West, A. P., Jr. *Science* **1992**, *255*, 1683.
- (4) Harris, R. A.; Tinoco, I., Jr. *J. Chem. Phys.* **1994**, *101*, 9289.
- (5) Li, L.; He, T.; Chen, D.; Wang, X.; Liu, F.-C. *J. Phys. Chem.* **1998**, *102*, 10385.
- (6) Jaszuński, M.; Rizzo, A. *Mol. Phys.* **1999**, *96*, 855.
- (7) Romero, R. H.; Vaara, J. *Chem. Phys. Lett.* **2004**, *400*, 226.
- (8) Ikäläinen, S.; Lantto, P.; Manninen, P.; Vaara, J. *J. Chem. Phys.* **2008**, *129*, 124102.
- (9) Savukov, I. M.; Lee, S. K.; Romalis, M. V. *Nature (London)* **2006**, *442*, 1021.
- (10) Lu, T.; He, M.; Chen, D.; He, T.; Liu, F.-C. *Chem. Phys. Lett.* **2009**, *479*, 14.
- (11) Ikäläinen, S.; Romalis, M. V.; Lantto, P.; Vaara, J. *Phys. Rev. Lett.* **2010**, *105*, 153001.
- (12) Pagliero, D.; Dong, W.; Sakellariou, D.; Meriles, C. A. *J. Chem. Phys.* **2010**, *133*, 154505.
- (13) Barron, L. D. *Molecular Light Scattering and Optical Activity*, 2nd ed.; Cambridge University Press: Cambridge, U.K., 2004; pp 145–147.
- (14) Pennanen, T. S.; Ikäläinen, S.; Lantto, P.; Vaara, J. Submitted for publication.
- (15) Manninen, P.; Vaara, J. *Comput. Chem.* **2006**, *27*, 434.
- (16) Dyall, K. G.; Fægri, K., Jr. *Introduction to Relativistic Quantum Chemistry*; Oxford University Press: New York, 2007; p 200.
- (17) Iliáš, M.; Saue, T. *J. Chem. Phys.* **2007**, *126*, 064102.
- (18) Ikäläinen, S.; Lantto, P.; Manninen, P.; Vaara, J. *Phys. Chem. Chem. Phys.* **2009**, *11*, 11404.
- (19) Vaara, J.; Pyykkö, P. *J. Chem. Phys.* **2003**, *118*, 2973.
- (20) Saue, T. *Adv. Quantum Chem.* **2005**, *48*, 383.
- (21) Hanni, M.; Lantto, P.; Iliáš, M.; Jensen, H. J. Aa.; Vaara, J. *J. Chem. Phys.* **2007**, *127*, 164313.
- (22) Kolb, D.; Johnson, W. R.; Shorer, P. *Phys. Rev. A* **1982**, *26*, 19.
- (23) Buckingham, A. D.; Stephens, P. J. *Annu. Rev. Phys. Chem.* **1966**, *17*, 399.
- (24) Buckingham, A. D. *Phil. Trans. R. Soc. London, Ser. A* **1979**, *293*, 239.
- (25) We use SI units throughout the paper. The positive Φ in eq 1 corresponds to a right-handed optical rotation as seen from the source.
- (26) Olsen, J.; Jørgensen, P. *J. Chem. Phys.* **1985**, *82*, 3235.
- (27) Chong, D. P. *Can. J. Chem.* **1995**, *73*, 79.
- (28) Kruunuhaka basis set tool kit, written by Manninen, P.; Lehtola, J. Release 2.0 (2011). <http://www.chem.helsinki.fi/~manninen/kruunuhaka/> (accessed February, 2011).
- (29) Dyall, K. G. *Theor. Chem. Acc.* **2006**, *115*, 441. Basis sets available from the Dirac Web site: <http://dirac.chem.sdu.dk> (accessed November, 2010).
- (30) DIRAC, a relativistic ab initio electronic structure program, release DIRAC10 (2010), written by Saue, T.; Visscher, L.; Jensen, H. J. Aa. with contributions from Bast, R.; Dyall, K. G.; Ekström, U.; Eliav, E.; Enevoldsen, T.; Fleig, T.; Gomes, A. S. P.; Henriksson, J.; Iliáš, M.; Jacob, Ch. R.; Knecht, S.; Nataraj, H. S.; Norman, P.; Olsen, J.; Pernpointner, M.; Ruud, K.; Schimmelpfennig, B.; Sikkema, J.; Thorvaldsen, A.; Thyssen, J.; Villaume, S.; Yamamoto, S. See <http://dirac.chem.vu.nl> (accessed November, 2010).
- (31) DALTON, a molecular electronic structure program, Release 2.0 (2005). See <http://www.kjemi.uio.no/software/dalton/dalton.html> (accessed November, 2010).
- (32) *CRC Handbook of Chemistry and Physics*, 82nd ed.; CRC Press: Boca Raton, FL, 2002; p 6–45.
- (33) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.
- (34) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
- (35) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
- (36) Stephens, P.; Devlin, F.; Chabalowski, C.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623.
- (37) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372.
- (38) Visscher, L.; Enevoldsen, T.; Saue, T.; Jensen, H. J. Aa.; Oddershede, J. *J. Comput. Chem.* **1999**, *20*, 1262.
- (39) Hettrema, H.; Jensen, H. J. Aa.; Jørgensen, P.; Olsen, J. *J. Chem. Phys.* **1992**, *97*, 1174.
- (40) Salek, P.; Vahtras, O.; Helgaker, T.; Ågren, H. *J. Chem. Phys.* **2002**, *117*, 9630.
- (41) Hättig, C.; Christiansen, O.; Koch, H.; Jørgensen, P. *Chem. Phys. Lett.* **1997**, *269*, 428.
- (42) Norman, P.; Jensen, H. J. Aa. *J. Chem. Phys.* **2004**, *121*, 6145.
- (43) Henriksson, J.; Saue, T.; Norman, P. *J. Chem. Phys.* **2008**, *128*, 024105.

(44) Dyall, K. G. *Theor. Chem. Acc.* **1998**, *99*, 366. Addendum *Theor. Chem. Acc.* **2002**, *108*, 365. Revision *Theor. Chem. Acc.* **2006**, *115*, 441. Basis sets available from the Dirac Web site: <http://dirac.chem.sdu.dk> (accessed Nov. 2011).

(45) Dyall, K. G. *Theor. Chem. Acc.* **2002**, *108*, 335. Erratum *Theor. Chem. Acc.* **2003**, *109*, 284. Revision *Theor. Chem. Acc.* **2006**, *115*, 441. Basis sets available from the Dirac Web site: <http://dirac.chem.sdu.dk> (accessed Nov. 2011).

(46) Bast, R.; Saue, T.; Henriksson, J.; Norman, P. J. *Chem. Phys.* **2009**, *130*, 024109.

(47) Ekström, U.; Norman, P.; Rizzo, A. J. *Chem. Phys.* **2005**, *122*, 074321.

(48) Ingersoll, L. R.; Liebenberg, D. H. *J. Opt. Soc. Am.* **1956**, *46*, 538.

(49) Aziz, R. A.; Slaman, M. J. *Mol. Phys.* **1986**, *57*, 825.

Intercalation of Transition Metals into Stacked Benzene Rings: A Model Study of the Intercalation of Transition Metals into Bilayered Graphene

Il Seung Youn,[†] Dong Young Kim,[†] N. Jiten Singh,[†] Sung Woo Park,[†] Jihee Youn,[†] and Kwang S. Kim^{*,†}

[†]Department of Chemistry and Department of Physics, Pohang University of Science and Technology, Pohang 790-784, Korea

S Supporting Information

ABSTRACT: Structures of neutral metal–dibenzene complexes, $M(C_6H_6)_2$ ($M = Sc–Zn$), are investigated by using Møller–Plesset second order perturbation theory (MP2). The benzene molecules change their conformation and shape upon complexation with the transition metals. We find two types of structures: (i) stacked forms for early transition metal complexes and (ii) distorted forms for late transition metal ones. The benzene molecules and the metal atom are bound together by δ bonds which originate from the interaction of π -MOs and d orbitals. The binding energy shows a maximum for $Cr(C_6H_6)_2$, which obeys the 18-electron rule. It is noticeable that $Mn(C_6H_6)_2$, a 19-electron complex, manages to have a stacked structure with an excess electron delocalized. For other late transition metal complexes having more than 19 electrons, the benzene molecules are bent or stray away from each other to reduce the electron density around a metal atom. For the early transition metals, the $M(C_6H_6)$ complexes are found to be more weakly bound than $M(C_6H_6)_2$. This is because the $M(C_6H_6)$ complexes do not have enough electrons to satisfy the 18-electron rule, and so the $M(C_6H_6)_2$ complexes generally tend to have tighter binding with a shorter benzene–metal length than the $M(C_6H_6)$ complexes, which is quite unusual. The present results could provide a possible explanation of why on the Ni surface graphene tends to grow in a few layers, while on the Cu surface the weak interaction between the copper surface and graphene allows for the formation of a single layer of graphene, in agreement with chemical vapor deposition experiments.

Bis(η^6 -benzene)chromium, $Cr(C_6H_6)_2$, is an 18-electron closed-shell compound including two benzene rings with a chromium atom at the center, which is one of the most well-known examples of organometallic sandwich complexes. Since its discovery by Fischer and Hafner,¹ numerous experimental and theoretical studies have been carried out to investigate how two benzene rings and a chromium atom interact and what kind of structure the complex forms.^{2–13} $Cr(C_6H_6)_2$ has the two eclipsed stacked forms of two benzene rings with the chromium atom placed at the midpoint of the two benzene centroids. On the basis of these studies, researchers have investigated electronic properties of the $Cr(C_6H_6)_2$ complex and the related cation complexes for a possible use as spin trap device or for the extension to carbon nanotubes and graphene.^{14–37} In addition, other similar molecules with transition metals have been studied for the same purpose,^{38–52} and the analogs such as graphene–metal hybrid materials have been utilized for electronic devices, biosensors, and the removal of hazardous materials.^{53–56} Nevertheless, the structure of the complexes of transition metals has not been properly studied at the high level of theory yet.

For $Cr(C_6H_6)_2$, the π character of each benzene interacts with d orbitals in the chromium atom; π -molecular orbitals (MOs) of benzene (Bz) molecules interact with the $d(xy)$ and $d(x^2 - y^2)$ orbitals in the chromium atom, forming the δ bond. In this case, the π – π interaction^{57–68} between two benzene molecules is very small because of their large separation, while the metal– π interaction^{69–74} between a metal atom and benzene molecules is dominant. Here, we investigated the structures of

bis(benzene)–first-row transition metal complexes (Bz_2M ; $M = Sc–Zn$) using ab initio calculations. Even though there have been many theoretical studies of Bz_2M , they used only density functional theory (DFT) methods,^{2,11,12,23,26,28,39,41,48,51} which have not been well tested for the interactions between benzene and central metal atoms. Thus, we have carried out MP2 calculations using the aug-cc-pVDZ (aVDZ) and aug-cc-pVTZ (aVTZ) basis sets. Since the highest occupied molecular orbitals (HOMO) are doubly degenerate, the Bz_2M complexes maintain uniformly stacked structures for $M = Sc–Cr$. Those structures support the 18-electron rule in organometallics. Even though Bz_2Mn has 19 electrons, it has the same structure with Bz_2Cr due to the nature of the HOMO, which diffuses the extra electron. For Bz_2M where $M = Fe–Zn$, the complexes cannot have well-ordered stacking forms because of the instability caused by too many excess electrons. Hence, two benzene molecules stray from each other or one benzene molecule is bent, changing the electron donation type from η^6 to η^4 or η^2 .

In order to compare the structural properties of the sandwiched complexes (Bz_2M) with those of the corresponding complexes having only one benzene molecule (BzM), we also examined their different natures in molecular bonding character. The critical interaction in BzM complexes is the one between π -MOs of benzene and $d(xz)$ or $d(yz)$ orbitals of the metal atom.^{10,33} This type of σ -bonding itself is stronger than δ bonding; yet, the BzM

Received: September 20, 2011

Published: November 21, 2011

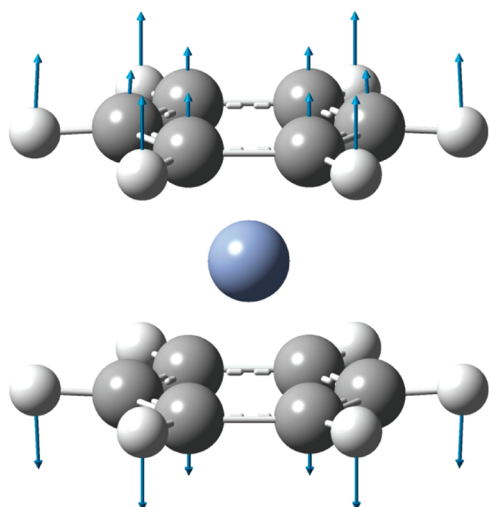


Figure 1. Breathing mode in Bz_2M .

complexes do not satisfy the 18-electron requirement because a ligand, benzene, donates electrons from only one side. On the other hand, two benzene molecules donate electrons, and then π -MOs interact with a metal atom from both above and below, satisfying the 18-electron requirement with stronger binding in Bz_2M . This results in a shorter distance between the metal atom and the benzene centroid (d_{M-Bz}) in the Bz_2M complexes than in the corresponding BzM ones for $M = Sc-Mn$, except for Bz_2Ti and Bz_2Cr .

We performed ab initio calculations using the Gaussian 09 package.⁷⁵ To search for low-lying energy structures, we have dealt with several probable structures based on the structure of $Cr(C_6H_6)_2$ with point group D_{6h} . In order to find the proper spin multiplicity of each metal atom in each complex, we optimized all structures with symmetry adaptation at each defined spin multiplicity by using MP2 with the aug-cc-pVDZ (aVDZ) basis set for carbon and hydrogen atoms and the CRENL effective core potentials (ECP)⁷⁶ for transition metal atoms. The frequency analysis was done to confirm the minimum-energy structures. The structures were reoptimized with basis set superposition error (BSSE) correction. The single point energy calculations were performed at the MP2 level of theory with the aug-cc-pVTZ (aVTZ) and CRENL ECP basis sets. The complete basis set limit energies^{77,78} were not made because of possible errors arising from large BSSEs at the aVDZ level. We studied natural bonding orbital (NBO) charges, binding energies (negative value of the interaction energies: $-\Delta E$), distances between the benzene and metal atom, and frequencies of a breathing mode (Figure 1).

DETERMINATION OF THE MOST STABLE SPIN CONFIGURATIONS

The possible spin multiplicities of each metal complex are 2 and 4 for Sc and Co; 1, 3, and 5 for Ti and Fe; 2, 4, and 6 for V and Mn; 1, 3, 5, and 7 for Cr; 1 and 3 for Ni; 2 for Cu; and 1 for Zn in the complexes. In the Bz_2M case, each complex has the most stable structure in the lowest spin multiplicity. The singlet and the triplet Fe complexes have similar energies. The spin multiplicity of each BzM complex at the lowest energy is dependent on the kind of metal atom: Sc, V, and Fe prefer 4, 4,

Table 1. MP2/aVTZ Results for the BzM Complexes

metal	point group	spin multiplicity	d_{M-Bz} (Å)	ΔE (kcal mol ⁻¹)	NBO charge of metals (a.u.) ^a
Sc	C_{2v}	4	1.923	-82.3	0.875
Ti	C_{2v}	1	1.604	-84.4	1.060
V	C_1	4	1.712	-37.4	0.996
Cr	C_{3v}	1	1.511	-208.7	0.697
Mn	C_1	2	1.887	-48.2	0.772
	C_1	4	1.610	-54.8	1.375
Fe	C_1	3	1.512	-30.8	1.239
Co	C_{2v}	2	1.503	-84.9	0.525
Ni	C_{3v}	1	1.441	-103.2	0.778
Cu	C_1	2	3.305	-1.9	-0.080
Zn	C_s	1	3.559	-2.0	-0.045

^aNBO charges are calculated at the MP2/aVDZ level.

and 3, respectively, and the others prefer the lowest spin multiplicities (see Supporting Information).

STRUCTURES OF BzM COMPLEXES

In the BzM complexes, geometry optimization was performed for several different spin multiplicities. These data are summarized in Table 1 (MP2/aVTZ) and in Table S1 (MP2/aVDZ) in the Supporting Information. Several spin states of BzM ($M = Sc, Ti, Cr, Mn, \text{ and } Co$) show attractive interactions. The lowest spin multiplicities exhibit stronger binding for all BzM 's except for $M = Sc, V, Mn, \text{ and } Fe$; for $M = Sc, V, \text{ and } Fe$, each spin multiplicity of 4, 4, and 3 shows stronger binding. The spin multiplicity of 2 in $BzMn$ shows the strongest binding at the MP2/aVDZ level, but the spin multiplicity of 4 does at the MP2/aVTZ level. As shown in Figure 2, the structures of BzM complexes are based on the structure of point group C_{6v} . Only the BzV has a bent benzene molecule below the metal atom. Not only the shape of benzene but also the d_{M-Bz} 's differ from each other.

The primary interactions, which give bonding character to a complex, are benzene π orbitals with metal $d(yz)$, $d(xz)$, and $d(z^2)$ orbitals (Figure 3b 3 and 4). Other important interactions are ones between π^* orbitals (top orbitals in Figure 3a) and the rest, two d orbitals. While those interactions lead to a bonding property, one between the π orbital shown in the bottom of Figure 3a and an s orbital brings out antibonding character. The key point is that there are stabilization and destabilization of some orbitals during the formation of the molecular orbitals (MOs). Five d orbitals locate differently depending on a metal atom and its spin state. In the $BzSc$ case, a scandium atom with a spin multiplicity of 4 has occupied frontier $d(yz)$, $d(z^2)$, and s orbitals, and other vacant 3d orbitals. Thus, when they form the $BzSc$ complex, high-lying $d(xy)$ and $d(x^2 - y^2)$ orbitals interact with π^* orbitals, which are in similar energy level, stabilizing the complex. A similar effect is caused from the formation of 3. On the other hand, the most stable π orbital is destabilized by interacting with the s orbital, forming the antibonding highest occupied molecular orbital (HOMO), 1. This antibonding HOMO causes relatively long distance d_{M-Bz} , 1.92 Å. The $BzTi$ complex resembles the situation, but there is not a large advantage to forming 3 or forming 1 as the HOMO. This results in similar binding energies for $BzSc$ and $BzTi$, but shorter d_{M-Bz} in $BzTi$, 1.60 Å. For BzV , a big energy loss comes from the

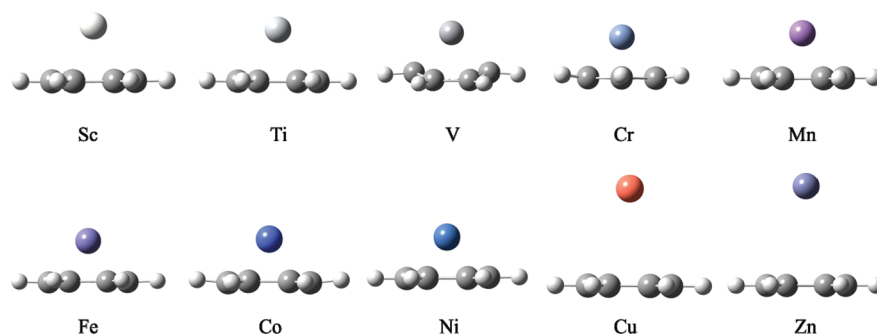


Figure 2. MP2/aVTZ predicted structures of Bz₂M complexes.

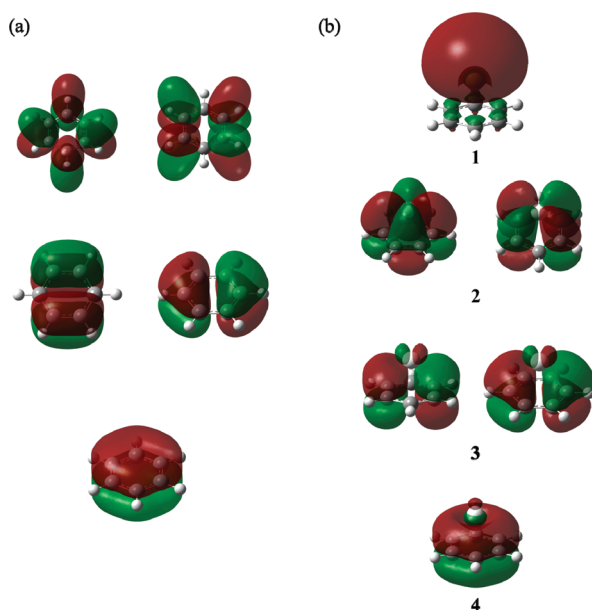


Figure 3. π orbitals of a benzene molecule (a) and important orbitals in Bz₂M complexes (b).

formation of 3, which includes destabilization of $d(yz)$ and $d(zx)$ orbitals, leading to a small binding energy. All 2 and 3 orbitals with stabilization of π^* , $d(xz)$, $d(yz)$, and $d(x^2 - y^2)$ orbitals are occupied in BzCr, and thus a benzene molecule and a chromium atom bind strongly. Some cases such as BzMn and BzFe cannot form bonding 2 or 3 orbitals due to a big difference in energy level between π^* and d orbitals, or they fail to gain big stabilization due to the same reason in spite of the formation of 2 and 3. A similar effect occurs intensively in BzCu and BzZn because their valence orbitals are fully occupied, so that extra electrons from a benzene molecule give rise to repulsion. Despite a big energy level difference, the Co and the Ni complexes form 2 and 3, which give great stabilization of the π^* orbitals and thus stronger binding energies than other BzM ($M = \text{Mn, Fe, Cu, and Zn}$).

The above analysis explains why the d_{M-Bz} is small for $M = \text{Sc-Ni}$ (1.4–1.9 Å), while the d_{M-Bz} is large for $M = \text{Cu and Zn}$ (3.305 and 3.559 Å, respectively). The small d_{M-Bz} 's are from the bonding 2 and 3, but the large ones are from the electron repulsion and the antibonding 1. This indicates that on the Ni surface graphene tends to grow in a few layers, while on the Cu surface the weak interaction between the copper surface and graphene would lead to the formation of a single layer of

graphene, in agreement with chemical vapor deposition (CVD) experiments.^{78–80} In addition, this significant difference in π –metal interactions between different metal atoms could be useful for ion sensing, such as conductance measurement through carbon-based electrodes such as graphene nanoribbons.^{81–83}

STRUCTURES OF Bz₂M COMPLEXES

In Bz₂M complexes, except for Bz₂Sc of point group C_2 , the symmetry is broken in all complexes. Nevertheless, the structures of early transition metal complexes (Bz₂Sc–Bz₂Mn) are based on D_{6h} -like structures. On the other hand, the late transition metal complexes (Bz₂Fe–Bz₂Zn) have structures in which two distorted benzene rings stray away from each other (Figure 4). Notable points in structures of the Bz₂M complexes are the d_{M-Bz} and the shape and arrangement of benzene molecules in each complex; while the Bz₂M complexes of early transition metals have their benzene molecules intact, as in the corresponding BzM complexes, a benzene molecule in the late transition metal complexes stray away from each other or one of them is severely distorted, as compared to the corresponding BzM complexes (Figures 2 and 4). This is due to the well-known 18-electron rule, which indicates that the number of electrons from the ligands and the metal atom may be summed up toward 18 to form a stable metal complex. The ligands are two benzene molecules here, and the metal atoms are first row transition metals. In each Bz₂M complex, one benzene ring donates six π electrons, and a metal atom has d electrons. Thus, the total number of electrons contributing to the bonding characters between two benzene rings and a metal is $6 \times 2 + d$ electrons. Table 2 gives MP2/aVTZ results for the Bz₂M complexes.

This analysis implies that in the range from Bz₂Sc to Bz₂Cr, d electrons occupying bonding orbitals lead to strong interactions. These bonding orbitals consist of π orbitals of the benzene molecules and $d(xy)$ and $d(x^2 - y^2)$ orbitals of the metal, resulting in δ bonding orbitals, as shown in Figure 5. This explains why the Bz₂Cr complex has the largest binding energy. This explanation is confirmed by comparing the calculation results of Bz₂V[−], Bz₂Cr[−], and Bz₂Mn⁺; the MP2/aVDZ and MP2/aVTZ results show that Bz₂V[−] (200 kcal mol^{−1} and not converged) and Bz₂Mn⁺ (242 kcal mol^{−1} and 241 kcal mol^{−1}), which are isoelectronic to Bz₂Cr (323 kcal mol^{−1} and 342 kcal mol^{−1}), have larger binding energies than Bz₂V (173 kcal mol^{−1} and 193 kcal mol^{−1}) and Bz₂Mn (181 kcal mol^{−1} and 167 kcal mol^{−1}), respectively, while Bz₂Cr[−] (200 kcal mol^{−1} and 213 kcal mol^{−1}) has a smaller binding energy than its neutral form.

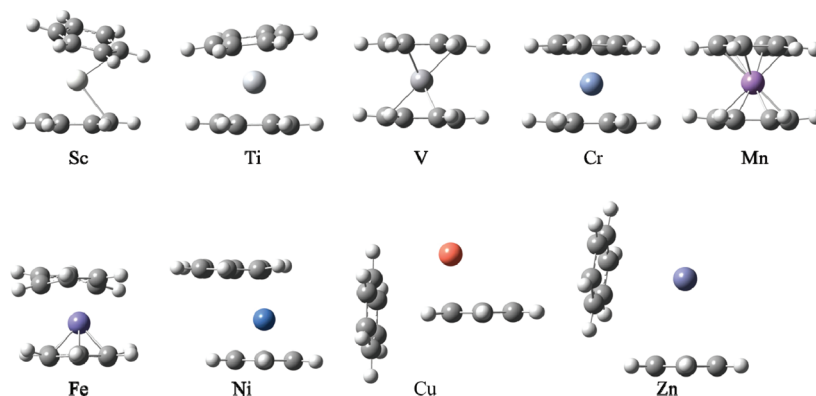


Figure 4. MP2/aVTZ predicted structures of the Bz_2M complexes.

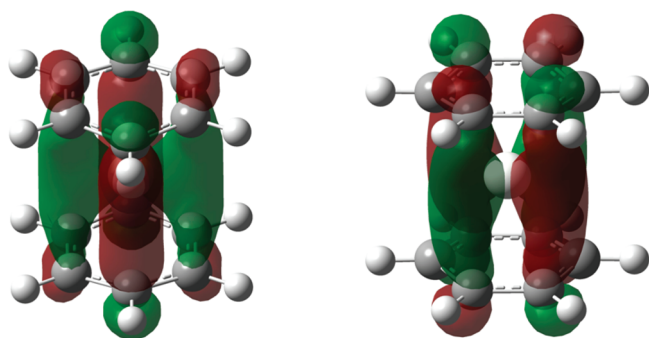


Figure 5. δ bonding orbitals in Bz_2M complexes with point group D_{6h} .

According to the MO analysis, it is expected that the d_{M-Bz} is shorter in the Bz_2M than in the corresponding BzM for $M = Sc-Cr$. The Bz_2Ti and Bz_2Cr , however, show slightly longer d_{M-Bz} in the Bz_2M . In the Bz_2Ti , all 2 and 3 orbitals are formed and are fully occupied, while δ bonding orbitals in Bz_2Ti are only partially occupied. Thus, even if Bz_2Ti shows stronger binding due to benzene stacking, it cannot obtain the full advantage of a decrease in d_{M-Bz} . It is also expected that the d_{M-Bz} in Bz_2M is the shortest in Bz_2V and Bz_2Cr due to the fact that 17 and 18 electrons fully occupy the orbitals in Figure 5. However, the d_{M-Bz} of Bz_2Cr is slightly longer than that of Bz_2V , even though the HOMO is a δ bonding orbital. This slight deviation from the d_{M-Bz} tendency may arise from the slightly negative charge accumulated on the chromium atom due to its large electron affinity (65 kJ mol^{-1}) as compared with the smaller electron affinity of V (51 kJ mol^{-1} ; Table 3). This negative charge of the metal repels the negative charges on carbon atoms in benzene molecules.

On the other hand, despite a 19-electron environment, the d_{M-Bz} of Bz_2Mn is the shortest among the five complexes, and the extent of the decrease for $BzMn$ is also large. Of course, the structure itself is less stable than any Bz_2M of early transition metals, except for Bz_2Sc , based on the small binding energy of $167 \text{ kcal mol}^{-1}$ for Bz_2Mn . As shown in Figure 6, a peculiar shape of the HOMO of Bz_2Mn , however, can dissipate electrons out of the central atom, and the complex is able to mitigate the electron repulsions. In fact, the positive NBO charge on the Mn atom in Table 3 supports the notion that Bz_2Mn dissipates electrons effectively (the electron affinity of Mn is $\sim 0 \text{ kJ mol}^{-1}$).

The intercalation of $FeCl_3$ inside the bilayer has recently been used for device fabrication.⁸⁴ It would be an interesting issue

Table 2. MP2/aVTZ Results for the Bz_2M Complexes^a

metal	spin multiplicity	d_{M-Bz} ($d_{Bz_2M} - d_{BzM}$) (Å)	ΔE (kcal mol^{-1})	$\Delta\Delta E$ (kcal mol^{-1}) ^b	freq. (cm^{-1}) ^c
Sc	2	1.913 (-0.010)	-98.2	+66.4	234
Ti	1	1.732 (+0.128)	-191.0	-22.2	262
V	2	1.532 (-0.181)	-192.6	-117.8	395
Cr	1	1.589 (+0.078)	-341.7	+75.7	285
Mn	2	1.473 (-0.137)	-167.2	-57.6	337
Fe	1	1.377 (-0.135), 2.049 ^d	-234.8	-173.1	<i>e</i>
Ni	1	1.743 (+0.302), 2.251 ^d	-132.7	+73.7	<i>e</i>
Cu	2	2.437 ^f (-0.868), 3.672	-9.8	-6.0	<i>e</i>
Zn	1	3.462, ^g 3.448 ^g (-0.111)	-7.0	-3.0	<i>e</i>

^a Bz_2Co was not optimized due to the convergence problem. ^b Cooperative binding energy difference: $\Delta\Delta E = \Delta E(Bz_2M) - 2 \times \Delta E(BzM)$ ^c Frequencies were calculated at the MP2/aVDZ level without BSSE correction. ^d Average distance between a metal and two nearest carbon atoms of the upper benzene. ^e The breathing mode is not defined. ^f Average distance between a metal and two nearest carbon atoms of each benzene. ^g Distance between Zn and each benzene centroid.

Table 3. d_{M-Bz} 's, Atomic Radii and NBO Charges of Metal in Bz_2M Complexes with Point Group of D_{6h} and BzM Complexes ($M = Sc - Mn$).^a

metal	atomic radius of metal atoms (Å)	d_{M-Bz} ($d_{Bz_2M} - d_{BzM}$) (Å)	Bz_2M	BzM
Sc	2.090	1.913 (-0.010)	1.121	0.875
Ti	2.000	1.732 (+0.128)	0.829	1.060
V	1.920	1.532 (-0.180)	0.216	0.996
Cr	1.850	1.589 (+0.078)	-0.076	0.697
Mn	1.790	1.473 (-0.137)	0.015	1.375

^a NBO charge of metal atoms (a.u.).

whether a single transition metal layer could be obtained inside bilayer graphene. In this regard, the metal-dibenzene structures could give interesting information for intercalated metal inside bilayered graphene.

The complexes of the late transition metals would be highly unstable if they maintain D_{6h} -like structure because in this case the number of electrons is larger than 18. Hence, the structures need to be distorted; one of benzene molecules strays so as not to donate all 6 electrons, donating fewer electrons to the central

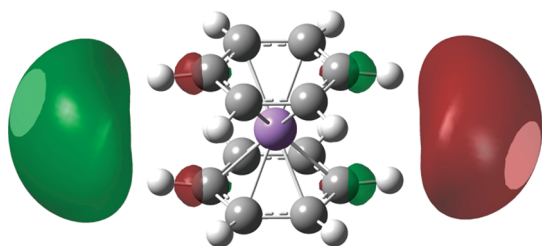


Figure 6. HOMO of Bz_2Mn with point group D_{6h} .

atom. Indeed, in the case of Bz_2Cu , it is interesting to note that since the interaction between Cu and Bz is very weak, the π -H interaction^{85,86} is dominant between two benzene molecules, and the structure is no longer a stacked form.

DISPERSION INTERACTION FOR $BzCu$ AND $BzZn$ COMPLEXES

Tables 1 and 2 present that Cu and Zn complexes show very weak binding (1.9 – 9.8 kcal mol⁻¹) in both BzM and Bz_2M complexes. The origin of these weak binding energies can be deduced from the NBO charges of metals. The NBO charges of Cu and Zn are -0.08 and -0.05 , respectively, indicating that there is no charge transfer from a metal atom to a benzene molecule in each complex. Hence, these weak binding energies of $BzCu$ and $BzZn$ are mainly due to the dispersion interaction. Note that the dispersion interaction is overestimated at the MP2 level of theory. To clarify this problem, we further performed the calculation at the level of coupled cluster theory with the inclusion of single and double excitations and perturbative inclusion of triple excitations (CCSD(T)) with aVDZ basis set using the Molpro package⁸⁷ for BzM ($M = Cu$ and Zn). The binding energies with BSSE correction are 1.3 kcal mol⁻¹ for $BzCu$ and 1.0 kcal mol⁻¹ for $BzZn$ at the level of CCSD(T)/aVDZ. Note that the MP2/aVTZ calculation results give 1.9 and 2.0 kcal mol⁻¹ for $BzCu$ and $BzZn$, respectively. Hence, in these cases, the MP2 level of theory gives a slightly overestimated dispersion interaction in comparison with the CCSD(T) level of theory.

In summary, we carried out a systematic study of Bz_2M complexes as compared with the corresponding BzM complexes. The results show sandwich structures for early transition metal complexes, while such sandwich structures are broken for late transition metal ones. The d_{M-Bz} in Bz_2M with doubly degenerate δ bonding orbitals tends to decrease from $M = Sc$ to $M = V$. It is quite interesting that even though the second coordination generally gives a longer coordination distance with smaller coordination energy than the first coordination, the present second coordination gives a shorter coordination distance with a larger coordination energy for the early transition metals because of the 18-electron rule. Unlike the Bz_2Cr , which gives some exception due to the negative NBO charge of Cr, the Bz_2Mn results in decreased d_{M-Bz} because of the diffuse HOMO. Structures of Bz_2M for late transition metals are distorted to avoid the instability caused by too many electrons around the central metal atom. As one of the two benzene molecules donates fewer electrons to a central atom, the whole structure is better stabilized. The present results provide a possible explanation of why graphene tends to grow in a few layers on the Ni surface, while on the Cu surface the weak interaction between the copper

surface and graphene allows for the formation of a single layer of graphene, in agreement with CVD experiments.

ASSOCIATED CONTENT

S Supporting Information. Discussion on the MP2/aVDZ results for BzM and Bz_2M . This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: kim@postech.ac.kr.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

This work was supported by NRF (National Honor Scientist Program, 2010-0020414; WCU, R32-2008-000-10180-0) and KISTI (KSC-2011-G3-02).

REFERENCES

- (1) Fischer, E. O.; Hafner, W. Z. *Naturforsch.* **1955**, *10b*, 665.
- (2) Rayón, D.; Frenking, G. *Organometallics* **2003**, *22*, 3304–3308.
- (3) Haaland, A. *Acta Chem. Scand.* **1965**, *19*, 41–46.
- (4) Albrecht, G.; Förster, E.; Sippel, D.; Eichkorn, F.; Kurras, E. Z. *Chem.* **1968**, *8*, 311.
- (5) Ngai, L. H.; Stafford, F. E.; Schäfer, L. *J. Am. Chem. Soc.* **1969**, *91*, 48–49.
- (6) Bochmann, M. *Organometallics 2, Complexes with TM-Carbon π bond*; Oxford Science Publication: Oxford, U.K., 1994.
- (7) Choi, K.-W.; Choi, S.; Sun, J. B.; Kim, S. K. *J. Chem. Phys.* **2007**, *126*, 034308.
- (8) Xiang, H.; Yang, J.; Hou, J. G.; Zhu, Q. *J. Am. Chem. Soc.* **2006**, *128*, 2310–2314.
- (9) Aspley, C. J.; Boxwell, C.; Buil, M. L.; Higgitt, C. L.; Long, C.; Perutz, R. N. *Chem. Commun.* **1999**, *11*, 1027–1028.
- (10) Muetterties, E. L.; Bleeke, J. R.; Wucherer, E. J. *Chem. Rev.* **1982**, *82*, 499–525.
- (11) Rayane, D.; Allouche, A.-R.; Antoine, R.; Broyer, M.; Compagnon, I.; Dugourd, P. *Chem. Phys. Lett.* **2003**, *375*, 506–510.
- (12) Yasuike, T.; Yabushita, S. *J. Phys. Chem. A.* **1999**, *103*, 4533–4542.
- (13) Jones, R. H.; Doerr, L. H.; Teat, S. J.; Wilson, C. C. *Chem. Phys. Lett.* **2000**, *319*, 423–426.
- (14) Ketkov, S. Y.; Selzle, H. L.; Schlag, E. W.; Domrachev, G. A. *Chem. Phys. Lett.* **2003**, *373*, 486–491.
- (15) Samuel, E.; Caurant, D.; Gourier, D.; Elschenbroich, Ch.; Agbaria, K. *J. Am. Chem. Soc.* **1998**, *120*, 8088–8092.
- (16) Calucci, L.; Cloke, F. G. N.; Englert, U.; Hitchcock, P. B.; Pampaloni, G.; Pinzino, C.; Puccinid, F.; Volpe, M. *Dalton Trans.* **2006**, 4228–4234.
- (17) Ketkov, S. Y.; Green, J. C.; Mehnert, C. P. *J. Chem. Soc., Faraday Trans.* **1997**, *93*, 2461–2466.
- (18) Xiang, H.; Yang, J.; Hou, J. G.; Zhu, Q. *J. Am. Chem. Soc.* **2006**, *128*, 2310–2314.
- (19) Choi, K.-W.; Ahn, D.-S.; Lee, S.; Kim, S. K. *J. Phys. Chem. A.* **2004**, *108*, 11292–11295.
- (20) Choi, K.-W.; Choi, S.; Ahn, D.-S.; Han, S.; Kang, T. Y.; Baek, S. J.; Kim, S. K. *J. Phys. Chem. A.* **2008**, *112*, 7125–7127.
- (21) Choi, K.-W.; Choi, S.; Baek, S. J.; Kim, S. K. *J. Chem. Phys.* **2007**, *126*, 034308.

- (22) Han, S.; Singh, N. J.; Kang, T. Y.; Choi, K.-W.; Choi, S.; Baek, S. J.; Kim, K. S.; Kim, S. K. *Phys. Chem. Chem. Phys.* **2010**, *12*, 7648–7653.
- (23) Sahnoun, R.; Mijoule, C. *J. Phys. Chem. A* **2001**, *105*, 6176–6181.
- (24) Li, Y.; Baer, T. *J. Phys. Chem. A* **2002**, *106*, 9820–9826.
- (25) Yi, H.-B.; Lee, H. M.; Kim, K. S. *J. Chem. Theory Comput.* **2009**, *5*, 1709–1717.
- (26) Lyssenko, K. A.; Korlyukov, A. A.; Golovanov, D. G.; Ketkov, S. Y.; Antipin, M. Y. *J. Phys. Chem. A* **2006**, *110*, 6545–6551.
- (27) Sohnlein, B. R.; Yang, D.-S. *J. Chem. Phys.* **2006**, *124*, 134305.
- (28) Bérces, A.; Ziegler, T. *J. Phys. Chem.* **1994**, *98*, 13233–13242.
- (29) Kim, D.; Hu, S.; Tarakeshwar, P.; Kim, K. S.; Lisy, J. M. *J. Phys. Chem. A* **2003**, *107*, 1228–1238.
- (30) Meyer, F.; Khan, F. A.; Armentrou, P. B. *J. Am. Chem. Soc.* **1995**, *117*, 9740–9748.
- (31) Singh, N. J.; Min, S. K.; Kim, D. Y.; Kim, K. S. *J. Chem. Theory Comput.* **2009**, *5*, 515–529.
- (32) Perrier, A.; Gourier, D.; Joubert, L.; Adamo, C. *Phys. Chem. Chem. Phys.* **2003**, *5*, 1337–1343.
- (33) Yi, H.; Diefenbach, M.; Choi, Y. C.; Lee, E. C.; Lee, H. M.; Hong, B. H.; Kim, K. S. *Chem.—Eur. J.* **2006**, *12*, 4885–4892.
- (34) Sohnlein, B. R.; Lei, Y.; Yang, D.-S. *J. Chem. Phys.* **2007**, *127*, 114302.
- (35) Philpott, M. R.; Kawazoe, Y. *Chem. Phys.* **2007**, *342*, 223–235.
- (36) Philpott, M. R.; Kawazoe, Y. *Chem. Phys.* **2008**, *348*, 69–82.
- (37) Singh, A. K.; Kumar, V.; Kawazoe, Y. *Eur. Phys. J. D* **2005**, *34*, 295–298.
- (38) Miyajima, K.; Nakajima, A.; Yabushita, S.; Knickelbein, M. B.; Kaya, K. *J. Am. Chem. Soc.* **2004**, *126*, 13202–13203.
- (39) Kandalam, A. K.; Rao, B. K.; Jena, P.; Pandey, R. *J. Chem. Phys.* **2004**, *120*, 10414–10422.
- (40) Béchamp, K.; Levesque, M.; Joly, H.; Manceron, L. *J. Phys. Chem. A* **2006**, *110*, 6023–6031.
- (41) Kua, J.; Tomlin, K. M. *J. Phys. Chem. A* **2006**, *110*, 11988–11994.
- (42) Philpott, M. R.; Kawazoe, Y. *J. Phys. Chem. A* **2008**, *112*, 2034–2042.
- (43) Kim, W. Y.; Kim, K. S. *Acc. Chem. Res.* **2010**, *43*, 111–120.
- (44) Kandalam, A. K.; Rao, B. K.; Jena, P.; Pandey, R. *J. Chem. Phys.* **2004**, *120*, 10414–10422.
- (45) Lee, E. C.; Choi, Y. C.; Kim, W. Y.; Singh, N. J.; Lee, S.; Shim, J. H.; Kim, K. S. *Chem.—Eur. J.* **2010**, *16*, 12141–12146.
- (46) Cho, Y.; Min, S. K.; Kim, W. Y.; Kim, K. S. *Phys. Chem. Chem. Phys.* **2011**, *115*, 6019–6023.
- (47) Kim, W. Y.; Choi, Y. C.; Min, S. K.; Cho, Y.; Kim, K. S. *Chem. Soc. Rev.* **2009**, *38*, 2319–2333.
- (48) Muhida, R.; Diño, W. A.; Rahman, M. M.; Kasai, H.; Nakanishi, H. *J. Phys. Soc. Jpn.* **2004**, *73*, 2292–2295.
- (49) Sceats, E. L.; Green, J. C. *Phys. Rev. B* **2007**, *75*, 245441.
- (50) Cho, W. J.; Cho, Y.; Min, S. K.; Kim, W. Y.; Kim, K. S. *J. Am. Chem. Soc.* **2011**, *133*, 9364–9369.
- (51) Rao, B. K.; Jena, P. *J. Chem. Phys.* **2002**, *116*, 1343–1349.
- (52) Valencia, H.; Gil, A.; Frapper, G. *J. Phys. Chem. C* **2010**, *114*, 14141–14153.
- (53) Myung, S.; Park, J.; Lee, H.; Kim, K. S.; Hong, S. *Adv. Mater.* **2010**, *22*, 2045–2049.
- (54) Myung, S.; Solanki, A.; Kim, C.; Park, J.; Kim, K. S.; Lee, K.-B. *Adv. Mater.* **2011**, *23*, 2221–2225.
- (55) Chandra, V.; Park, J.; Chun, Y.; Lee, J. W.; Hwang, I.-C.; Kim, K. S. *ACS Nano* **2010**, *4*, 3979–3986.
- (56) Chandra, V.; Kim, K. S. *Chem. Commun.* **2011**, *47*, 3942–3944.
- (57) Riley, K. E.; Pitok, M.; Jurečka, P.; Hobza, P. *Chem. Rev.* **2010**, *110*, 5023–5063.
- (58) Lee, E. C.; Kim, D.; Jurečka, P.; Tarakeshwar, P.; Hobza, P.; Kim, K. S. *J. Phys. Chem. A* **2007**, *111*, 3446–3457.
- (59) Sinnokrot, M. O.; Sherrill, C. D. *J. Phys. Chem. A* **2006**, *110*, 10656–10688.
- (60) Singh, N. J.; Min, S. K.; Kim, D. Y.; Kim, K. S. *J. Chem. Theory Comput.* **2009**, *5*, 515–529.
- (61) Lee, J. Y.; Hong, B. H.; Kim, W. Y.; Min, S. K.; Kim, Y.; Jouravlev, M. V.; Bose, R.; Kim, K. S.; Hwang, I.-C.; Kaufman, L. J.; Wong, C. W.; Kim, P.; Kim, K. S. *Nature* **2009**, *460*, 498–501.
- (62) Kim, K. S.; Tarakeshwar, P.; Lee, J. Y. *Chem. Rev.* **2000**, *100*, 4145–4185.
- (63) Ringer, A. L.; Sinnokrot, M. O.; Lively, R. P.; Sherrill, C. D. *Chem.—Eur. J.* **2006**, *12*, 3821–3828.
- (64) Špirko, V.; Hobza, P. *Chem. Phys. Chem.* **2006**, *7*, 640–643.
- (65) Timoshkin, A. Y.; Frenking, G. *Inorg. Chem.* **2003**, *42*, 60–69.
- (66) Sinnokrot, M. O.; Sherrill, C. D. *J. Am. Chem. Soc.* **2004**, *126*, 7690–7697.
- (67) Pitoňák, M.; Neogrády, P.; Řezáč, J.; Jurečka, P.; Urban, M.; Hobza, P. *J. Chem. Theory Comput.* **2008**, *4*, 1829–1834.
- (68) Sinnokrot, M. O.; Sherrill, C. D. *J. Phys. Chem. A* **2003**, *107*, 8377–8379.
- (69) Kim, D.; Hu, S.; Tarakeshwar, P.; Kim, K. S.; Lisy, J. M. *J. Phys. Chem. A* **2003**, *107*, 1228–1238.
- (70) Ihm, H.; Yun, S.; Kim, H. G.; Kim, J. K.; Kim, K. S. *Org. Lett.* **2002**, *4*, 2897–2900.
- (71) Yi, H.-B.; Lee, H. M.; Kim, K. S. *J. Chem. Theory Comput.* **2009**, *5*, 1709–1717.
- (72) Yi, H.-B.; Diefenbach, M.; Choi, Y. C.; Lee, E. C.; Lee, H. M.; Hong, B. H.; Kim, K. S. *Chem.—Eur. J.* **2006**, *12*, 4885–4892.
- (73) Rayón, V. M.; Frenking, G. *Chem.—Eur. J.* **2002**, *8*, 4693–4707.
- (74) Frunzke, J.; Lein, M.; Frenking, G. *Organometallics* **2002**, *21*, 3351–3359.
- (75) 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, V.; 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 09*, Revision A.02; Gaussian, Inc.: Wallingford, CT, 2009.
- (76) Hurley, M. M.; Pacios, L. F.; Christiansen, P. A. *J. Chem. Phys.* **1986**, *84*, 6840–6853.
- (77) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *J. Chem. Phys.* **1997**, *106*, 9639–9646.
- (78) Min, S. K.; Lee, E. C.; Lee, H. M.; Kim, D. Y.; Kim, D.; Kim, K. S. *J. Comput. Chem.* **2008**, *29*, 1208–1221.
- (79) Kim, K. S.; Zhao, Y.; Jang, H.; Lee, S. Y.; Kim, J. M.; Kim, K. S.; Ahn, J. -H.; Kim, P.; Choi, J.-H.; Hong, B. H. *Nature* **2009**, *457*, 706–710.
- (80) Li, X.; Cai, W.; An, J.; Kim, S.; Nah, J.; Yang, D.; Piner, R.; Velamakanni, A.; Jung, I.; Tutuc, E.; Banerjee, S. K.; Colombo, L.; Ruoff, R. S. *Science* **2009**, *324*, 1312–1314.
- (81) Bae, S.; Kim, H.; Lee, Y.; Xu, X.; Park, J.-S.; Zheng, Y.; Balakrishnan, J.; Lei, T.; Kim, H. R.; Song, Y. I.; Kim, Y.-J.; Kim, K. S.; Özyilmaz, B.; Ahn, J.-H.; Hong, B. H.; Iijima, S. *Nat. Nanotechnol.* **2010**, *8*, 574–578.
- (82) Min, S. K.; Kim, W. Y.; Cho, Y.; Kim, K. S. *Nat. Nanotechnol.* **2011**, *6*, 162–165.
- (83) Kim, W. Y.; Kim, K. S. *Nat. Nanotechnol.* **2008**, *3*, 408–412.
- (84) Kim, N.; Kim, K. S.; Jung, N.; Brus, L.; Kim, P. *Nano Lett.* **2011**, *11*, 860–865.
- (85) Lee, E. C.; Hong, B. H.; Lee, J. Y.; Kim, J. C.; Kim, D.; Kim, Y.; Tarakeshwar, P.; Kim, K. S. *J. Am. Chem. Soc.* **2005**, *127*, 4530–4537.

(86) Tarakeshwar, P.; Choi, H. S.; Kim, K. S. *J. Am. Chem. Soc.* **2001**, *123*, 3323–3331.

(87) Werner, H.-J.; Knowles, P. J.; Manby, F. R.; Schütz, M.; Celani, P.; Knizia, G.; Korona, T.; Lindh, R.; 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.; Hesselmann, A.; Hetzer, G.; Hrenar, T.; Jansen, G.; Köppl, C.; Liu, Y.; Lloyd, A. W.; Mata, R. A.; May, A. J.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pflüger, K.; Pitzer, R.; Reiher, M.; Shiozaki, T.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T.; Wang, M.; Wolf, A. *MOLPRO*, version 2010.1; University College Cardiff Consultants Limited: Wales, U.K., 2010. <http://www.molpro.net> (accessed Nov. 2011).

Eigensystem Representation of the Electronic Susceptibility Tensor for Intermolecular Interactions within Density Functional Theory

A. Scherrer, V. Verschinin, and D. Sebastiani*

Dahlem Center for Complex Quantum Systems, Physics Department, Free University Berlin, Arnimallee 14, 14195 Berlin, Germany

ABSTRACT: We present an efficient implementation of the electronic susceptibility tensor within density functional theory. The susceptibility is represented by means of its eigensystem, which is computed using an iterative Lanczos diagonalization technique for the susceptibility tensor within density functional perturbation theory. We show that a representation in a finite basis of eigenstates is sufficiently accurate to compute the linear response of the electronic density to external potentials. Once the eigensystem representation is computed, the actual response computation can be done at very low computational cost. The method is applied to the water molecule in a dipole field as a benchmark system. The results illustrate the potential of the approach for the first-principles calculation of supramolecular interactions in complex disordered systems in the condensed phase.

1. INTRODUCTION

Most spectroscopic techniques in physics and chemistry measure the response of the investigated system to an external field, that is, an external modification of the environmental situation. More specifically, optical spectroscopies use an electric field, while magnetic resonance spectroscopies work on the basis of a magnetic field. In many cases, the response of the system is primarily of electronic nature, meaning that the electrons in the system change their quantum state. This change in state then emits or absorbs, for example, radiation, which in turn is measured by the experiment.

In many cases, the external field is small as compared to the typical electronic energy spectrum, in particular excitation energies. In such cases, it is possible to consider the external field as a small perturbation, and within the context of quantum mechanics, perturbation theory can be applied to determine the linear response of the electronic subsystem to the perturbation. In this framework, the induced change in the electronic orbitals is assumed to be proportional to the strength of the external field. Within the framework of first-principles electronic structure theories, in particular density functional theory (DFT^{1–3}), this linearity also applies to the electronic density, giving rise to density functional perturbation theory.^{4–7}

It is straightforward to show (see section 3) that the linear response of the electron density to a (nonimaginary) local perturbation $\hat{H}^{(1)}$ can be written as

$$n^{(1)}(\mathbf{r}) = \int d\mathbf{r}' \chi(\mathbf{r}, \mathbf{r}') H^{(1)}(\mathbf{r}') \quad (1)$$

with a universal linear response function $\chi(\mathbf{r}, \mathbf{r}')$ (independent of $\hat{H}^{(1)}$). χ is formally a tensor in a continuous basis and therefore difficult to handle in practice. In this work, we have developed an implementation similar to a very recently published approach for the related problem of the static dielectric matrix^{8–10} to approximate the susceptibility tensor by expansion in its eigensystem representation. The method uses repeated calculation of the response in eq 1 via density functional perturbation theory (DFPT).^{11,12}

It turns out that the spectrum of $\chi(\mathbf{r}, \mathbf{r}')$ converges sufficiently quickly to allow for an efficient representation of the tensor in its finite eigensystem representation. This assumption has already been validated by several recent studies,^{13–18} in which a similar scheme was used to compute the dielectric response matrix, the RPA correlation and self-energies, as well as optical spectra of condensed-phase systems.

It is known that, in addition to dielectric response properties, a perturbation theory-based Ansatz is in principle also capable of representing supramolecular interaction energies to a very high accuracy.^{19,20} In this Article, we develop the electronic linear response approach within density functional perturbation theory for subsequent application to a perturbative calculation of such intermolecular interactions.

Complementary to the existing implementation by Galli et al.,^{9,10,13} we specifically aim at computing interaction energies and atomic forces between the components of supramolecular systems, for example, complex liquids (water, ionic liquids) or molecular crystals. Our present implementation has not yet been optimized to tackle such systems in a black-box manner, but our results show that the approach yields highly accurate results. We believe that the method can be used to calculate ab initio level interaction energies at a very low computational cost.

2. THEORY

2.1. Susceptibility Tensor with DFPT. Within DFPT,^{6,11,12} all relevant quantum quantities (Hamiltonian, orbitals, density) are expanded expressed by their unperturbed and perturbed components, for example, $\hat{H} = \hat{H}^{(0)} + \lambda \hat{H}^{(1)}$. The first-order response of the orbitals can formally be calculated by

$$|\psi_i^{(1)}\rangle = -(\hat{H}^{(0)} - E_i^{(0)})^{-1} P_e \hat{H}^{(1)} |\psi_i^{(0)}\rangle \quad (2)$$

with $P_e = 1 - \sum_{j, \text{occ}} |\psi_j^{(0)}\rangle \langle \psi_j^{(0)}|$. Assuming that the perturbation

Received: September 30, 2011

Published: November 23, 2011

Hamiltonian $\hat{H}^{(1)}$ is local, the resulting density response is given by eq 1 with

$$\chi(\mathbf{r}, \mathbf{r}') = -\sum_{i=1}^N [\psi_i^{*(0)}(\mathbf{r}) \langle \mathbf{r} | (\hat{H}^{(0)} - E_i^{(0)})^{-1} P_e | \mathbf{r}' \rangle \psi_i^{(0)}(\mathbf{r}') + \text{cc}] \quad (3)$$

This expression of $\chi(\mathbf{r}, \mathbf{r}')$ does not provide a feasible way for its calculation because the dimensions of $\chi(\mathbf{r}, \mathbf{r}')$ are continuous and any suitable real-space discretization would yield matrices with dimensions too large for explicit matrix inversions. However, from eq 3 it is apparent that $\chi(\mathbf{r}, \mathbf{r}')$ is real and symmetric. Hence, $\hat{\chi}$ can be expressed on the basis of its eigenstates $|\chi_\xi\rangle$, defined via $\hat{\chi}|\chi_\xi\rangle = \chi_\xi|\chi_\xi\rangle$, as

$$\hat{\chi} = \sum_{\xi} |\chi_\xi\rangle \chi_\xi \langle \chi_\xi| \quad (4)$$

It is important to note that in this decomposition the contribution of the eigenstates is weighted by their eigenvalues χ_ξ . Eigenstates with zero eigenvalue do not contribute to the summation and can thus be omitted. For nonmetallic systems, the spectrum $\{\chi_\xi\}$ is bound from above, because the expression $(\hat{H}^{(0)} - E_i^{(0)})^{-1}$ in eq 2 is limited by the inverse of the HOMO–LUMO energy gap. Hence, if the spectrum decays sufficiently fast, most of the eigenvalues may be omitted to a good approximation; it is a valid approximation to omit most of the eigenstates.

$$\hat{\chi} \approx \sum_{\xi}^{N_{\max}} |\chi_\xi\rangle \chi_\xi \langle \chi_\xi| \quad (5)$$

With this at hand, the approximate determination of $\hat{\chi}$ turns into the problem of finding the eigenvectors with the corresponding largest eigenvalues.

2.2. Lanczos Diagonalization. For the calculation of the eigenvectors $|\chi_\xi\rangle$ corresponding to the largest eigenvalues χ_ξ , we resort to an iterative diagonalization scheme (Lanczos, see section 3). The Lanczos method is a Krylov-space approach, which requires the repeated application of the operator that shall be diagonalized to a given vector $|\mu\rangle$. For our electronic susceptibility tensor, one such application $|\nu\rangle = \hat{\chi}|\mu\rangle$ corresponds to solving the DFPT eq 2 once with $\mu = \hat{H}^{(1)}$ for $\nu = n^{(1)}$. This operation is hence straightforward and requires a computational effort similar to a ground-state total energy calculation.

2.3. Polarizability. One of the physical observables closely related to the electronic susceptibility tensor $\hat{\chi}$ is the electric polarizability tensor $\alpha = \text{dp}/d\mathbf{E}$, where \mathbf{E} is a homogeneous electric field acting as a perturbation. The induced polarization $\delta\mathbf{p}$ is given by

$$\delta\mathbf{p} = \int d\mathbf{r} \mathbf{r} n^{(1)}(\mathbf{r}) \quad (6)$$

$$= \int d\mathbf{r} \mathbf{r} \int d\mathbf{r}' \chi(\mathbf{r}, \mathbf{r}') e\mathbf{E}_0 \cdot \mathbf{r}' \quad (7)$$

Using the eigenstate representation of $\hat{\chi}$ and the first moments of its eigenstates $\beta_{\mu,\xi} = \int d\mathbf{r} \chi_\xi(\mathbf{r}) r_\mu$ gives

$$\alpha_{\mu\nu} = \sum_{\xi} \chi_\xi e \int d\mathbf{r} \chi_\xi(\mathbf{r}) r_\mu \int d\mathbf{r}' \chi_\xi(\mathbf{r}') r'_\nu \quad (8)$$

$$= \sum_{\xi} \chi_\xi e \beta_{\mu,\xi} \beta_{\nu,\xi} \quad (9)$$

Thus, the electric polarizability α can be computed directly from the susceptibility eigenfunctions and can therefore serve as a perfect tool measure for the convergence analysis for the finite expansion eq 5.

3. IMPLEMENTATION

3.1. Lanczos Algorithm. Formally, the numerical problem to solve is the determination of the eigenvectors with the largest eigenvalues of an unknown Hermitian matrix A with only its action on a vector available. This matrix A is not known explicitly, but only by its action on a given vector. In this work, this diagonalization task is done using the Hermitian Lanczos algorithm. It is an iterative method to obtain approximate eigenvectors, and the corresponding orthogonal projection $B_m \in \mathbb{C}^{m \times m}$ of a Hermitian matrix $A \in \mathbb{C}^{n \times n}$ with $m \ll n$ uses Krylov subspaces \mathcal{K}_m to iteratively create the needed subspace needed for the orthogonal projection.^{21,22}

The implemented version of the algorithm is

- Choose \mathbf{v}_1 with $|\mathbf{v}_1| = 1$. Set $\beta_1 = 0$, $\mathbf{v}_0 = 0$.
- Iterate for $j = 1, 2, \dots, m$

$$\tilde{\mathbf{w}}_j = A\mathbf{v}_j \quad (10)$$

$$\alpha_j = \tilde{\mathbf{w}}_j \cdot \mathbf{v}_j \quad (11)$$

$$\mathbf{w}_j = \hat{P} \tilde{\mathbf{w}}_j \quad (12)$$

$$\beta_{j+1} = |\mathbf{w}_j| \quad (13)$$

$$\mathbf{v}_{j+1} = \mathbf{w}_j / \beta_{j+1} \quad (14)$$

The orthonormalization \hat{P} in eq 12 is performed with respect to all vectors already found. With exact arithmetics, only the first two vectors would be sufficient.

The calculated vectors \mathbf{v}_j form an orthonormal basis $V_m = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m]$, and the resulting orthogonal projection matrix $B_m = V_m^\dagger A V_m$ has the desirable properties of being real, tridiagonal, and symmetric. The calculated coefficients α_j and β_j are its diagonal and off-diagonal elements, respectively. Because $m \ll n$, the diagonalization of the Rayleigh–Ritz procedure is numerically feasible and can be done with, for example, a QR-decomposition with scaling that scales as $\mathcal{O}(m^2)$. The resulting eigenvectors \mathbf{u}_j of B_m are called Ritz vectors and contain the coefficients for the approximate expansion of the original eigenvectors of the matrix A in the basis V_m . The approximate eigenvectors are calculated ordered by the absolute value of their corresponding eigenvalues as desired for their application in this context. The Lanczos algorithm yields the approximate eigenvectors in the order of decreasing eigenvalues. This means that the extremal part of the spectrum is obtained first, which fits the idea of the representation according to eq 5.

3.2. Underlying Electronic Hamiltonian. The concepts presented so far are general and independent of the electronic structure method chosen. In this work, we use DFT and obtain the response in eq 1 via density functional perturbation theory¹² in the implementation framework of the CPMD software.²³ We use separable norm-conserving pseudopotentials,²⁴ the BLYP

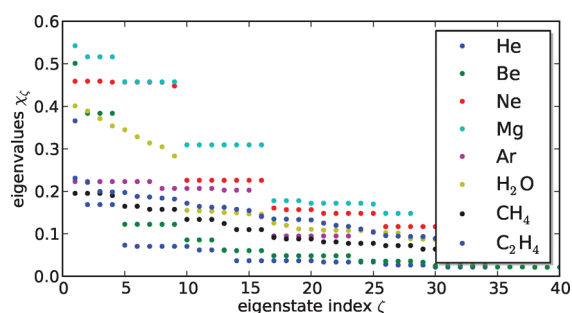


Figure 1. First eigenvalues of the lightest closed-shell atoms and the molecules H₂O, CH₄, and C₂H₄.

gradient corrections functionals, and a PW cutoff of 70 Ry with simple cubic system symmetry. The cell sizes are 7 Å for isolated atoms and 10 Å for molecular systems.

3.3. Convergence. Mathematically speaking, the choice of the initial vector determines the shape and the maximum dimension of the calculated Krylov subspaces, that is, also the number of possible iterations m . Ideally, the start vector should be a linear combination of all relevant eigenvectors, that is, all eigenvectors with significantly nonzero eigenvalue. In our particular case of the electronic susceptibility tensor, we have found that the electronic ground-state density shifted by q/Ω (with $q = \int_{\Omega} d\mathbf{r} \rho(\mathbf{r})$) is sufficient as an initial vector a good choice. However, the impact of nonexact arithmetics effectively increases the maximum dimension and reduces the sensitivity of the method to the initial conditions.

At any iteration of the Lanczos cycle in an actual calculation, only a part of the Ritz vectors u_j are numerically accurate eigenvectors of A . A suitable measure for the quality of a given u_j is the absolute value of its last element u_{jm} . This element represents the overlap of the true eigenvector with the latest calculated vector v_m . We have verified empirically that this convergence criterion is an excellent and reliable choice for our purposes.

We have set the convergence criterion for the eigenvectors to $u_{jm} \leq 10^{-5}$. After about 5000 Lanczos iterations, we find that typically one-half of the resulting Ritz vectors u_j can be considered as converged. This ratio remains approximately constant also for subsequent iterations.

4. RESULTS

To illustrate the validity and the versatility of the approach, we have applied our finite basis representation of the electric susceptibility tensor to several complementary molecular systems. Specifically, we have computed the following:

- the three-dimensional shape of the eigenstates and decay of the eigenvalues for isolated atoms of different elements (He, Ne, Ar, Be, Mg) and small molecules (H₂O, CH₄, C₂H₄);
- the polarizability of H₂O, CH₄, C₂H₄, and C₂H₆; and
- the induced polarization of Be, H₂O, CH₄, and C₂H₄ due to the inhomogeneous electric field of a salt ion pair Na⁺Cl⁻.

4.1. Eigenstates and Eigenvalues. To obtain an idea of the numerical shape of the spectrum of $\hat{\chi}$, we have plotted the initial eigenvalues for a set of isolated atoms and molecules in Figure 1.

For the single atoms, the convergence is stepwise with several eigenvalues of equal or similar size. The decay of the eigenvalue spectrum resembles that of the energy spectrum of Z/r -type Coulomb potentials. Such Z/r potentials yield degenerate states

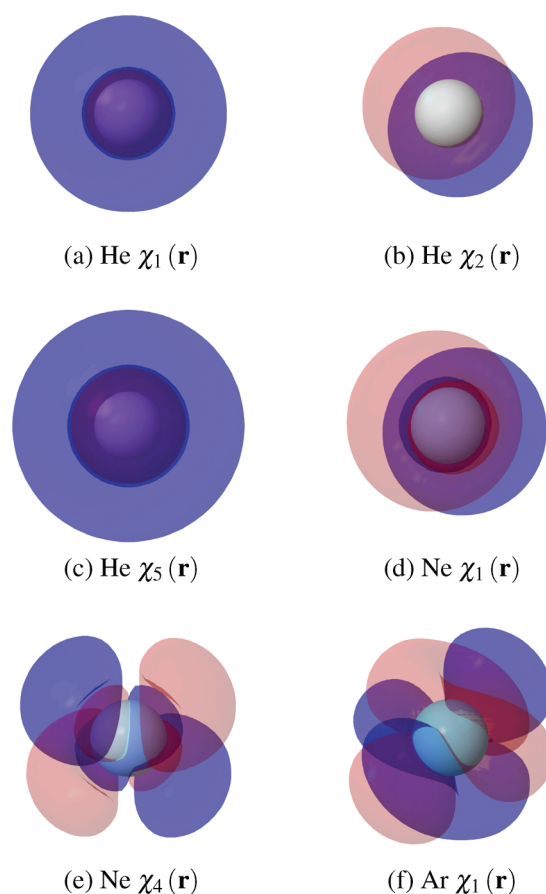


Figure 2. First eigenstates of He, Ne, and Ar. The isosurface plot is done with Jmol using cutoffs of +0.01 (red) and -0.01 (blue).

with characteristic symmetry properties. The eigenstates of the susceptibility tensor turn out to exhibit very similar symmetries, as depicted in Figure 2 for He, Ne, and Ar. The analogy to the more familiar orbital model is obvious.

The first eigenstate of helium is spherically symmetric and nondegenerate, whereas the 3-fold degeneracy of the second to fourth eigenstate of helium and the first three eigenstates of neon come along with a symmetry similar to p-orbitals of the hydrogen atom. The analogy may not be applied strictly as shows the example of argon, where the first eigenvalue is 7-fold degenerate, which has no trivial counterpart in the hydrogen case.

Naturally, the degeneracies observed for spherically symmetric atoms are lifted for less symmetric molecules such as H₂O or C_xH_y. The eigenvalues of the latter do not show exact degeneracies. A selection of the first eigenstates is plotted in Figure 3. The symmetries of the eigenstates are related to the symmetries of the molecule. Furthermore, the parity changes with respect to the different symmetry axes, which is relevant for the convergence properties in the following. Generally, the eigenstates exhibit different parity properties for the different molecular symmetry axes.

The eigenstates of higher order have more complex symmetries. As an example, the 30th eigenstates of Be and H₂O are shown in Figure 4. The overall tendency is an increase in the number of nodes, that is, more oscillations and steeper slopes. Again, this is analogous to the behavior of higher orbitals in the orbital model with higher quantum numbers.

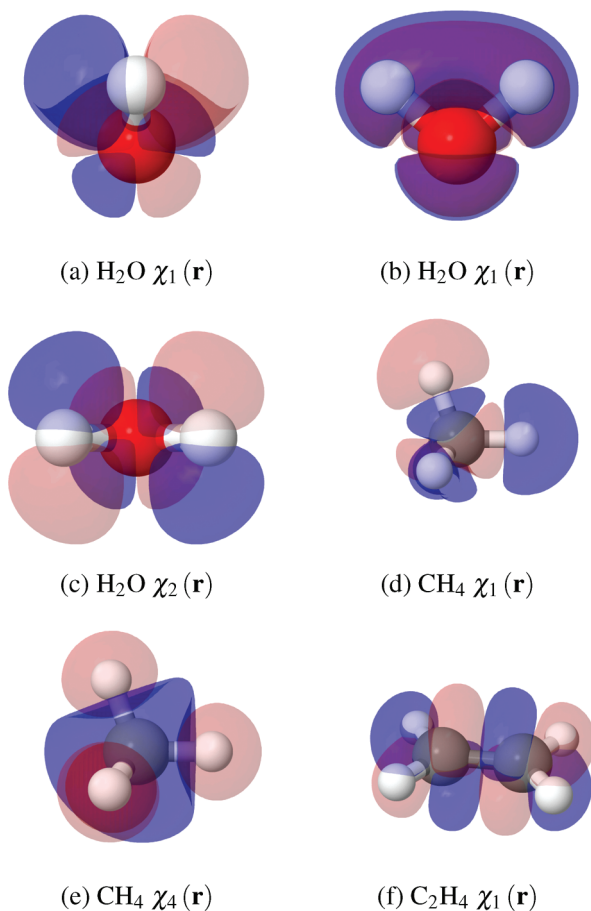


Figure 3. First eigenstates of H₂O, CH₄, and C₂H₄.

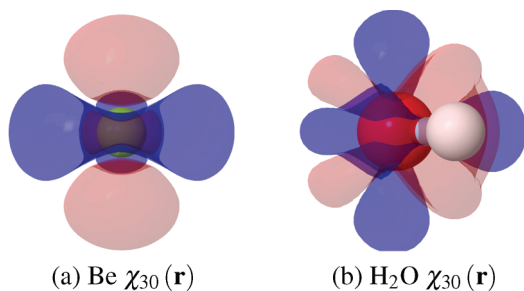


Figure 4. 30th eigenstate of Be and H₂O.

The quick decay of the spectrum of $\hat{\chi}$ is an important property that we observed in all considered systems. It allows the truncation of the summation according to eq 5 with a moderate value for N_{\max} . The asymptotic behavior of the eigenvalues is depicted in Figure 5 and shows an algebraic decline. The difference between Be and the molecules indicates that more complex systems converge more slowly. To a certain extent, the decay characteristics are determined by the total number of occupied orbitals. This effect has already been discussed by Galli.¹³

Because the eigenstates have the character of a density response, they must obey charge conservation. This means that $\int d\mathbf{r} \chi_\nu(\mathbf{r}) \stackrel{\pm}{=} 0$ for all eigenstates. We have checked numerically that this charge conservation property is fulfilled up to machine precision of 10^{-14} .

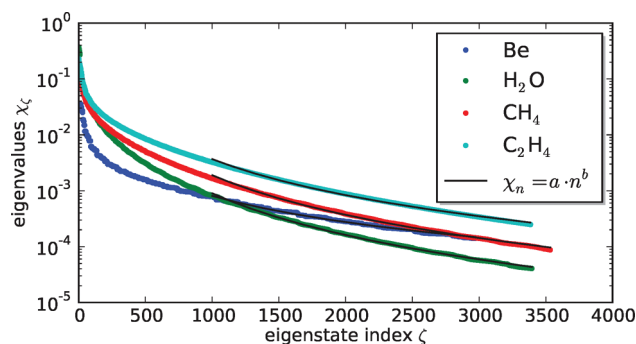


Figure 5. Asymptotic behavior of eigenvalues. The fitting parameters for the systems are $a = (3.7, 10.1, 10.0, 9.2)$ and $b = (-1.6, -2.5, -2.4, -2.1)$ for Be, H₂O, CH₄, and C₂H₄, respectively.

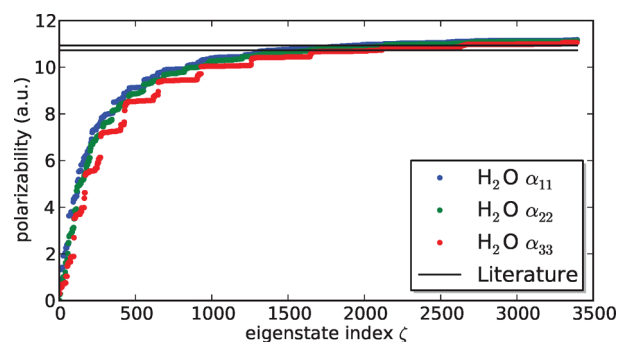


Figure 6. Main diagonal elements of the polarizability tensor of H₂O. The given literature values are for DFT calculations with GGA functional.^{12,25}

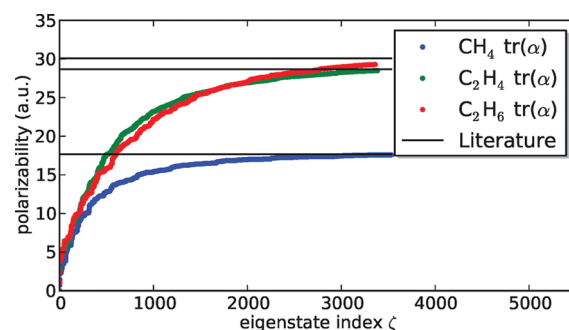


Figure 7. Polarizability of CH₄, C₂H₄, and C₂H₆. Literature values are as in Figure 6.

4.2. Polarizability. The polarizability in the presence of a homogeneous infinitesimal electric field can be computed with standard DFPT¹² and via eq 9. In Figure 6, we illustrate the numerical convergence of the susceptibility-based expression to the literature value from refs 12,25. The graph shows several plateaus, which indicates that many of the eigenstates have a vanishing contribution to the polarizability. This is due to the symmetry of the eigenstates with respect to the direction of the field. If the parity of χ_ν is even in a given direction, the corresponding β_ν vanishes.

As a further benchmark, we have computed the traces of the polarizability tensor $tr(\alpha)$ for the organic molecules CH₄, C₂H₄, and C₂H₆. The convergence as a function of the total number of

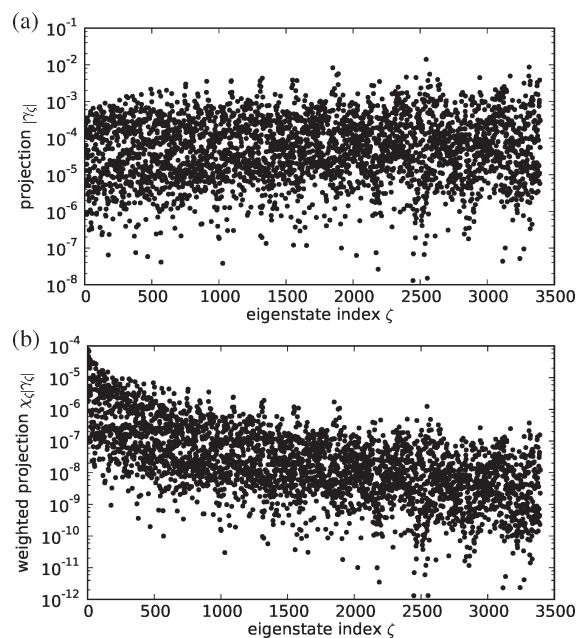


Figure 8. Projection coefficients and weighted projection coefficients of the perturbing potential on the eigenstates of H_2O .

eigenstates is depicted in Figure 7. Again, the convergence depends moderately on the number of electrons in the molecule.

Summarizing the results in this section, we conclude that the susceptibility in its low-rank approximation in eq 5 is able to represent the electronic polarizability tensor for our set of small molecules to a very good accuracy.

4.3. Polarization by Finite Charges. As a first step toward the description of intermolecular interaction, we have looked at the polarization of a water molecule in the inhomogeneous electrostatic field of a Na^+Cl^- ion pair. Denoting the potential of the two ions by $H^{(1)} := V_{[\text{Na}^+\text{Cl}^-]}^{\text{Coulomb}}$, the response of a system X (e.g., H_2O) requires the calculation of the projections $\chi_{\zeta, [X]}$ of the perturbing potential on the eigenstates of the system X :

$$\gamma_{\zeta, [X, \text{Na}^+\text{Cl}^-]} := \langle \chi_{\zeta, [X]} | V_{[\text{Na}^+\text{Cl}^-]}^{\text{Coulomb}} \rangle \quad (15)$$

Combining eqs 4, 1, and 15, the density response is given as

$$n_{[X, \text{Na}^+\text{Cl}^-]}^{(1)}(\mathbf{r}) = \sum_{\zeta} \chi_{\zeta, [X]}(\mathbf{r}) \gamma_{\zeta, [X, \text{Na}^+\text{Cl}^-]} \quad (16)$$

For a water molecule, the resulting coefficients $\gamma_{\zeta, [\text{H}_2\text{O}, \text{Na}^+\text{Cl}^-]}$ are depicted in Figure 8a. Up to 3000 eigenstates, no particular trend is observed. In fact, the coefficients $\gamma_{\zeta, [\text{H}_2\text{O}, \text{Na}^+\text{Cl}^-]}$ show a broad scattered distribution between 10^{-3} and 10^{-6} .

The contribution of a particular eigenstate in eq 16 is proportional to the product of the projection coefficient and its corresponding eigenvalue; this product is depicted in Figure 8b. While the width of the distribution of the weighted coefficients remains constant for the whole spectrum, the center of the distribution decreases for large eigenstate indices ζ . However, the decay is relatively slow, even after several thousand eigenstates. Together with the large width of the weighted projection band, this indicates that at the present stage of the implementation, there is still a large part of the computational effort spent on insignificant eigenstates. This problem is related to the mutual symmetry properties of eigenstates and the perturbation as in the case of the polarizability tensor.

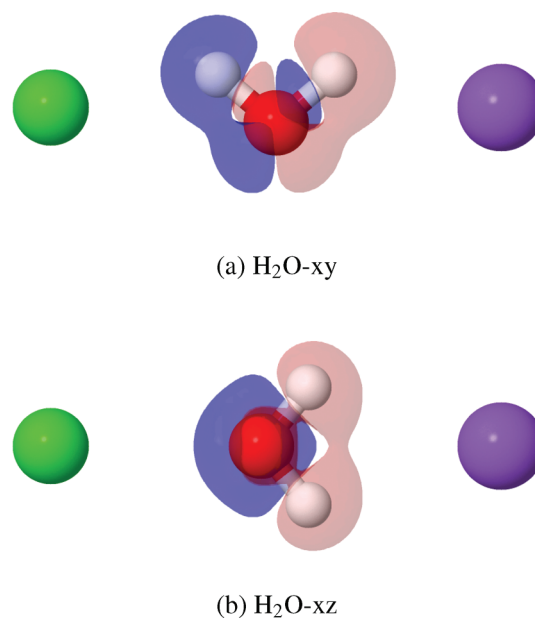


Figure 9. Linear response densities for H_2O for different orientations in the Na^+Cl^- potential, calculated via DFPT (left, Cl^- ; right, Na^+).

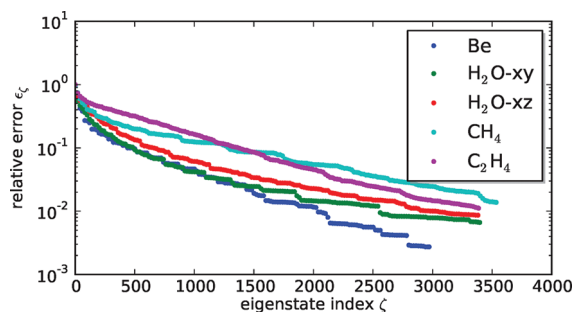


Figure 10. Relative error of the approximated $\hat{\chi}$ -linear response to the directly calculated one for Be, H_2O , CH_4 , and C_2H_4 .

We now want to compute the actual linear response density of our water molecule, that is, the charge displacement due to the field of the Na^+Cl^- atoms. This quantity is obtained by summing all data points in Figure 8b. This response density is illustrated in Figure 9, showing the shift of the electric cloud away from the Cl^- side toward the Na^+ ion. Interestingly, the opposite polarization is observed locally around the oxygen atom (Figure 9a). This is a purely quantum effect, which cannot be reproduced by a simple “polarizable charge distribution” model.

When using eq 16 to compute this charge displacement, the sum has to be truncated. To quantify the approximation error due to the truncation, we compare the result of eq 16 $n^{(1)}(\zeta)$ with the response calculated directly via DFPT n^{DFPT} in dependency of the truncation index ζ :

$$\epsilon(\zeta) = \frac{|n^{\text{DFPT}} - n^{(1)}(\zeta)|_2}{|n^{\text{DFPT}}|_2} \quad (17)$$

The asymptotic behavior of this relative error is depicted in Figure 10. The convergence is mainly smooth rippled. The results for the different orientations of H_2O show a slight difference, which is due to the different symmetry properties of the eigenstates with respect to the symmetry of the perturbing

potential. The convergence is similar for all molecules considered here. Only a slight dependence on the complexity of the system is observed. Depending on this complexity, a relative error of $\sim 1\%$ may be achieved from about 3000–4000 converged eigenstates.

5. CONCLUSION

We have used a finite representation of the electronic linear susceptibility tensor for the computationally efficient calculation of molecular interactions, in particular electrostatic polarization, within density functional theory. Our results show that it is possible to compute such polarization effects with arbitrary accuracy for the response density $\delta\rho(\mathbf{r})$, typically 1% when considering about 3000 eigenstates.

The calculation of the response due to an arbitrary perturbation requires only few vector multiplications and no actual wave function optimizations. This means that the approach may be able to achieve virtually exact DFT interaction energies at a fraction of the computational cost. The application of actual supramolecular systems still needs further implementation efforts, which are presently underway.

AUTHOR INFORMATION

Corresponding Author

*E-mail: daniel.sebastiani@fu-berlin.de.

Note

The authors declare no competing financial interest.

ACKNOWLEDGMENT

We are grateful to Michele Parrinello for valuable discussions, which initiated this project. This work was financially supported by the Deutsche Forschungsgemeinschaft under grants Se 1008/5 and Se 1008/6.

REFERENCES

- (1) Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, B864.
- (2) Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *140*, A1133.
- (3) Jones, R. O.; Gunnarsson, O. *Rev. Mod. Phys.* **1989**, *61*, 689–746.
- (4) Gonze, X. *Phys. Rev. A* **1995**, *52*, 1096.
- (5) Gonze, X.; Allan, D. C.; Teter, M. P. *Phys. Rev. Lett.* **1992**, *68*, 3603.
- (6) Gonze, X.; Vigneron, J. P. *Phys. Rev. B* **1989**, *39*, 13120.
- (7) Giannozzi, P.; de Gironcoli, S.; Pavone, P.; Baroni, S. *Phys. Rev. B* **1991**, *43*, 7231.
- (8) Hamel, S.; Williamson, A. J.; Wilson, H. F.; Gygi, F.; Galli, G.; Ratner, E.; Wack, D. *Appl. Phys. Lett.* **2008**, *92*, 043115.
- (9) Lu, D.; Gygi, F.; Galli, G. *Phys. Rev. Lett.* **2008**, *100*, 147601.
- (10) Wilson, H. F.; Gygi, F.; Galli, G. *Phys. Rev. B* **2008**, *78*, 113303.
- (11) Baroni, S.; de Gironcoli, S.; del Corso, A.; Giannozzi, P. *Rev. Mod. Phys.* **2001**, *73*, 515.
- (12) Putrino, A.; Sebastiani, D.; Parrinello, M. *J. Chem. Phys.* **2000**, *113*, 7102–7109.
- (13) Wilson, H. F.; Lu, D.; Gygi, F.; Galli, G. *Phys. Rev. B* **2009**, *79*, 245106.
- (14) Lu, D.; Nguyen, H.-V.; Galli, G. *J. Chem. Phys.* **2010**, *133*, 154110.
- (15) Pham, T. A.; Li, T.; Shankar, S.; Gygi, F.; Galli, G. *Appl. Phys. Lett.* **2010**, *96*, 062902.
- (16) Rocca, D.; Lu, D.; Galli, G. *J. Chem. Phys.* **2010**, *133*, 164109.
- (17) Pham, T. A.; Li, T.; Shankar, S.; Gygi, F.; Galli, G. *Phys. Rev. B* **2011**, *84*, 045308.
- (18) Kang, W.; Hybertsen, M. S. *Phys. Rev. B* **2010**, *82*, 195108.
- (19) Benoit, D.; Sebastiani, D.; Parrinello, M. *Phys. Rev. Lett.* **2001**, *87*, 226401.
- (20) Filippone, F.; Parrinello, M. *Chem. Phys. Lett.* **2001**, *345*, 179–182.
- (21) Lanczos, C. *J. Res. Natl. Bur. Stand.* **1951**, *45*, 255–282.
- (22) Saad, Y. *Numerical Methods for Large Nonsymmetric Eigenvalue Problems*, 1st ed.; Manchester University Press: Manchester, UK, 1992; pp 42–142.
- (23) Hutter, J.; et al. *Computer code CPMD, version 3.12.0*; Copyright IBM Corp. and MPI-FKF Stuttgart, 1990–2007; <http://www.cpmid.org>, accessed 4/1/2011.
- (24) Goedecker, S.; Teter, M.; Hutter, J. *Phys. Rev. B* **1996**, *54*, 1703–1710.
- (25) Porezag, D.; Pederson, M. R. *Phys. Rev. B* **1996**, *54*, 7830–7836.

Bathochromic Shift in Green Fluorescent Protein: A Puzzle for QM/MM Approaches

Claudia Filippi,^{*,†} Francesco Buda,^{*,‡} Leonardo Guidoni,[§] and Adalgisa Sinicropi^{*,||}

[†]Faculty of Science and Technology and MESA+ Institute of Nanotechnology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands

[‡]Leiden Institute of Chemistry, Leiden University, P.O. Box 9502, 2300 RA Leiden, The Netherlands

[§]Dipartimento di Chimica, Ingegneria Chimica e Materiali, Università dell'Aquila, Via Campo di Pile, 67100 L'Aquila, Italy

^{||}Dipartimento di Chimica, Università di Siena, Via A. De Gasperi 2, 53100 Siena, Italy

 Supporting Information

ABSTRACT: We present an extensive investigation of the vertical excitations of the anionic and neutral forms of wild-type green fluorescent protein using time-dependent density functional theory (TDDFT), multiconfigurational perturbation theory (CASPT2), and quantum Monte Carlo (QMC) methods within a quantum mechanics/molecular mechanics (QM/MM) scheme. The protein models are constructed via room-temperature QM/MM molecular dynamics simulations based on DFT and are representative of an average configuration of the chromophore–protein complex. We thoroughly verify the reliability of our structures through simulations with an extended QM region, different nonpolarizable force fields, as well as partial reoptimization with the CASPT2 approach. When computing the excitations, we find that wave function as well as density functional theory methods with long-range corrected functionals agree in the gas phase with the extrapolation of solution experiments but fail in reproducing the bathochromic shift in the protein, which should be particularly significant in the neutral case. In particular, while all methods correctly predict a shift in the absorption between the anionic and neutral forms of the protein, the location of the theoretical absorption maxima is significantly blue-shifted and too close to the gas-phase values. These results point to either an intrinsic limitation of nonpolarizable force-field embedding in the computation of the excitations or to the need to explore alternative protonation states of amino acids in the close vicinity of the chromophore.

1. INTRODUCTION

In previous decades, the development and application of intrinsically fluorescent proteins¹ has launched a revolution in molecular cell biology. Fluorescent proteins are routinely employed to dynamically visualize cellular processes in living organisms and are responsible for significant advances in fluorescence spectroscopy, enabling for instance bioimaging techniques with subdiffraction resolution.² Green fluorescent protein (GFP) is the prototype of this class of proteins and, together with its mutants, one of the most widely used fluorescent markers in cell biology.³ Because of its technological relevance, GFP is very well characterized experimentally and represents therefore a perfect playground to investigate the effectiveness of commonly used as well as novel quantum mechanics/molecular mechanics (QM/MM) approaches.

In this paper, we employ a variety of electronic structure approaches to compute the vertical excitations of the neutral A and anionic B forms of wild-type GFP. These are the protonation states of the chromophore responsible for the two room-temperature absorption peaks at 398 nm (3.12 eV) and 478 nm (2.59 eV), respectively.⁴ For these two forms, we construct the protein models via room-temperature QM/MM molecular dynamics simulations based on density functional theory (DFT) and thoroughly test their reliability by performing simulations with extended QM regions, different nonpolarizable force fields, and even partially relaxing the QM component with the multiconfigurational perturbation theory (CASPT2) approach.

The absorption spectra of the neutral and anionic forms are then computed with the use of time-dependent density functional theory (TDDFT), CASPT2, and quantum Monte Carlo (QMC) methods within a QM/MM scheme.

We find that, in the gas phase, wave function and density functional methods with long-range corrected functionals agree well with the extrapolation of solution experiments⁵ but, when introducing the protein environment, do not yield the desired bathochromic shift, which should be particularly significant in the neutral case. In particular, we obtain theoretical excitations between 2.8 and 3.2 eV for the B form and in the range of 3.4–3.6 eV for the A form. Therefore, while all methods predict a shift in the absorption between the anionic and neutral protein forms, the location of the theoretical absorption maxima is significantly blue-shifted and too close to the gas-phase values. Our extensive tests on the structures indicate that the source of the problem does not lie in the computational details of the construction within the standard QM/MM prescription. Moreover, since the theoretical techniques to compute the excitation spectrum appear to be reliable in the gas phase and display an overall agreement among each other in the protein, we also rule out that the origin of the problem lies in our choice of QM method to compute the excitation. Therefore, our results point at two possible sources of error still largely unexplored, namely, that

Received: October 4, 2011

Published: November 21, 2011

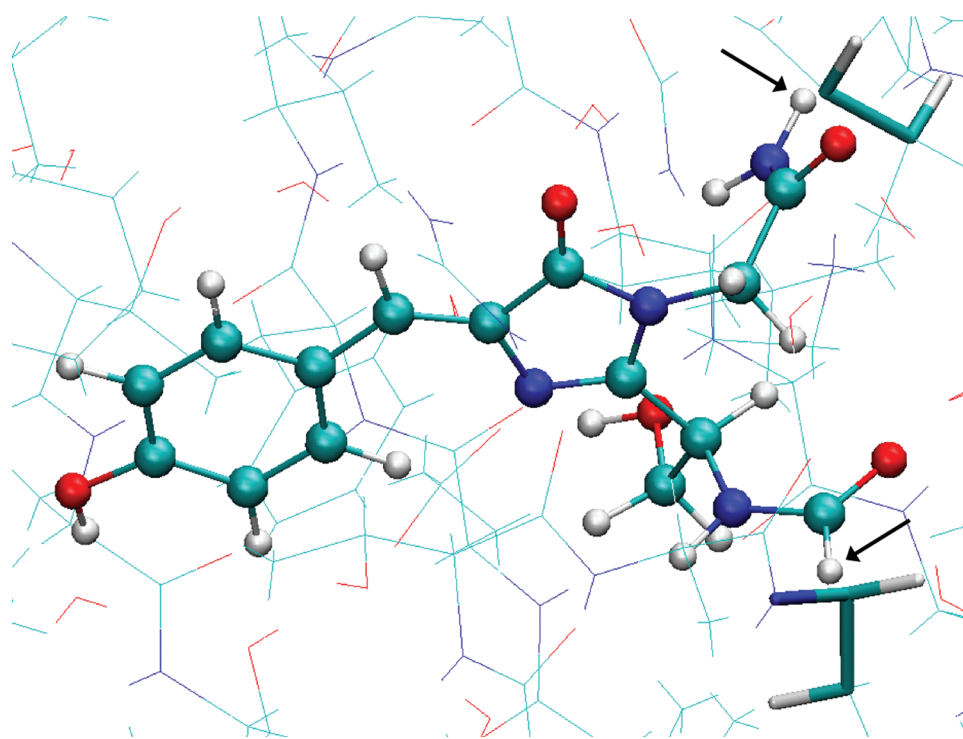


Figure 1. Partitioning between the QM and MM regions shown for the A form of wild-type GFP. The QM/MM cuts are at the C_{OOH}–C_α bond of Phe64 and the N–C_α bond of Val68. The arrows indicate the hydrogen-link atoms between the QM and the MM. The MM residues closest to the QM atoms are represented with cylinders. This image is produced with VMD.³⁷

some amino acids in the protein have a different protonation than commonly accepted or that the protein environment must be described beyond nonpolarizable force fields in the computation of the excitations.

In section 2, we describe the computational details and focus on the construction of the protein models for the neutral A and anionic B as well as alternative nonstandard forms. The results for the vertical excitations are presented in section 3. Finally, we discuss the structural models and the theoretical excitations in section 4 and conclude in section 5.

2. METHODS AND MODELS

2.1. Computational Details. The Amber suite of programs⁶ in combination with the Amber 03 force field⁷ and the TIP3P water model⁸ is used in the setup of the protein models and for the subsequent MM equilibrations.

The CPMD code⁹ is used for all ab initio molecular dynamics simulations described by quantum mechanics. The Gromos96 code¹⁰ with the Amber 03 force field is employed to describe MM atoms within the QM/MM approach.^{11,12} The Kohn–Sham orbitals are expanded in a plane-wave basis set with a kinetic energy cutoff of 70 Ry. We employ Martins–Troullier norm-conserving pseudopotentials¹³ and the Perdew, Becke, and Ernzerhof (PBE) generalized gradient approximation for the exchange–correlation functional.^{14,15} The size of the QM box is such that the distance between periodic replicas is about 8 Å in every direction. The annealing and the room-temperature Car–Parrinello molecular dynamics simulations are performed with a time step of 0.075 fs and a value of 400 au for the fictitious electronic mass in the Car–Parrinello Lagrangian.

To compute the excitation energies, we employ TDDFT, the complete active space self-consistent field (CASSCF) method with its perturbative extension (CASPT2), and QMC methods.

For the TDDFT and TDDFT/MM calculations, we use the Gaussian 09 code¹⁶ with default convergence parameters. The BLYP,^{17,18} B3LYP,¹⁹ CAM-B3LYP,²⁰ LC-BLYP,²¹ and the LC- ω PBE²² functionals are employed together with the aug-cc-pVDZ basis set. The effect of using the cc-pVDZ, cc-pVTZ, and aug-cc-pVTZ basis sets is also tested. In the TDDFT/MM calculations, we use the Amber 99 point charges for consistency with the CASPT2 calculations.

The complete active space calculations are performed using MOLCAS 7.2.²³ In the CASPT2 calculations, we employ the default IPEA zero-order Hamiltonian²⁴ unless otherwise stated and indicate if an additional imaginary constant level shift²⁵ is added to the Hamiltonian. In the CASPT2 calculations, we do not correlate as many of the lowest σ orbitals as there are heavy atoms in the molecule. For all models, we use the Cholesky decomposition of the two-electron integrals^{26,27} with the threshold of 10^{-8} . Default convergence criteria are used for all calculations. The CASPT2/MM calculations are performed with the MOLCAS 7.2 package coupled with a modified version of the MM package Tinker 4.2.²⁸ The Amber 99 force field²⁹ and TIP3P water model are employed.

The program package CHAMP³⁰ is used for the QMC calculations. We employ scalar-relativistic energy-consistent Hartree–Fock pseudopotentials³¹ where the carbon and nitrogen 1s electrons are replaced by a nonsingular s-nonlocal pseudopotential and the hydrogen potential is softened by removing the Coulomb divergence. We use the Gaussian basis sets³¹ specifically constructed for our pseudopotentials. In particular, we employ the cc-pVDZ basis augmented with diffuse s and p functions³² on the heavy atoms and denoted as D+. Different Jastrow factors are

used to describe the correlation with different atom types, and for each atom type, the Jastrow factor consists of an exponential of the sum of two fifth-order polynomials of the electron–nucleus and the electron–electron distances, respectively.³³ The determinant components of the ground and excited states are obtained in state-average CASSCF calculations performed with the program GAMESS(US).³⁴ The final CAS expansions are expressed on the CASSCF natural orbitals and may be truncated with an appropriate threshold on the configuration state functions (CSF) coefficients for use in the QMC calculations. The union of the surviving CSFs for the states of interest is kept in the final Jastrow–Slater wave functions. The Jastrow correlation factor and the CI coefficients are optimized by energy minimization in a state-averaged fashion within variational Monte Carlo (VMC).³⁵ The pseudopotentials are treated beyond the locality approximation,³⁶ and an imaginary time step of 0.04 or 0.06 au is used in the diffusion Monte Carlo (DMC) calculations. The QMC/MM calculations are performed using the electrostatic coupling scheme as in the QM/MM approach employed in the CPMD code. For the starting trial wave function, CASSCF/MM calculations are performed within GAMESS(US), and the divergence of point charges at the origin is removed to resemble the embedding scheme employed within CPMD.

2.2. Protein Models. For all forms of GFP, the protein models are initially equilibrated via MM simulations at room temperature. The hydrogens and water molecules are first equilibrated while constraining the positions of the heavy atoms. Subsequently, a MM isothermal and isobaric simulation is performed at 300 K and 1 atm, keeping only the chromophore coordinates fixed. Finally, the structure is refined in a simulated annealing run within QM/MM, where the chromophore is treated quantum mechanically and allowed to relax. The boundary between the QM and the MM regions is set through the single C_{OOH}–C_α bond of Phe64 and the single N–C_α bond of Val68, as shown in Figure 1. The possible protonation states of the QM chromophore model are shown in Figure 2.

2.2.1. The Neutral and Anionic Forms. The starting structure for the construction of the neutral A form is the X-ray structure³⁸ at 1.90 Å resolution (entry 1GFL in the Protein Data Bank³⁹). On the basis of the most likely hydrogen-bonding configuration, the histidine residues numbered 25, 148, 181, 199, and 217 are protonated at their δ nitrogen while the remaining histidine residues are protonated at their ϵ nitrogen. For the protonation of the glutamic acids, particular attention is given to Glu222, which is deprotonated in the A form and will be the proton acceptor in the proton shuffle between the neutral and the anionic forms.

To simulate solution conditions, the protein is placed at the center of a cubic MM simulation box surrounded by 12 Å of water molecules in each direction. Moreover, counterions are added to achieve physiologic conditions with a saline concentration of 0.15 M and to ensure that the cell is neutral. The total simulation box contains around 70 000 atoms. A short isothermal and isobaric MM equilibration at a fixed chromophore is performed with a time step of 0.5 fs until the temperature fluctuations are less than 5% and the density is constant within 2%. A quenched structure is then obtained via an energy minimization procedure. Starting from the structure obtained in the MM equilibration, we perform a short QM/MM run of about 1 ps at 300 K, followed by an annealing period of about 0.5 ps, where the velocities are rescaled at each step by a factor in the range of 0.99–0.999. The temperature of the final cooled system is less than 0.5 K.

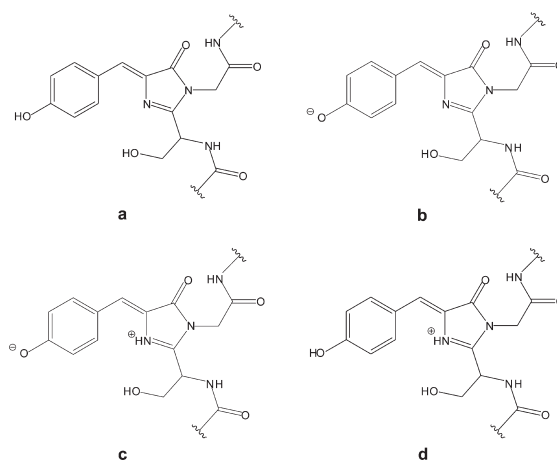


Figure 2. Possible protonation states of the QM chromophore model of GFP: (a) neutral, (b) anionic, (c) zwitterionic, and (d) cationic.

The anionic I form is obtained by deprotonating the chromophore of our equilibrated A form and adding the proton to Glu222. A QM/MM simulated annealing procedure is then performed via an isothermic molecular dynamics simulation followed by a velocity rescaling as described for the A form.

Since no crystal structure of the wild-type anionic B form is available, the model for this form is obtained starting from the crystallographic structure of a mutant stabilizing the anionic chromophore with an environment believed to be close to that of wild-type GFP in the B form. Following the theoretical work of Nifosi and Tozzini,⁴⁰ we start from the crystallographic structure⁴¹ of mutant S65T (code 1EMG in the Protein Data Bank). This mutant differs from the wild-type structure (1GFL) for the substitution of Ser65 with a threonine (Thr65). To restore the wild-type structure, we simply undo this mutation by reintroducing Ser65 and adjusting its orientation to form a hydrogen bond with Glu222, which is now protonated. Starting from this structure, we add 10 Å of water molecules in each direction and the appropriate number of counterions, leading to a system of about 31 000 atoms. Note that the number of atoms is approximately half that of the A form, as the starting crystal structures of the A and B forms are a dimer and a monomer, respectively. We then proceed with the classical MM equilibration and the subsequent QM/MM dynamics and annealing as described for the A form.

2.2.2. Nonstandard Forms. We also investigate the stability of alternative protonation states of the chromophore and surrounding environment via additional QM/MM simulations. We test the possible stability of a solvated hydronium in proximity of the chromophore following the experimental suggestions of ref 42 as well as the existence of the zwitterionic and cationic forms in the presence of a protonated and deprotonated Glu222 residue. In these simulations, the QM region is extended to include the anionic chromophore; the side chains of Thr203, Ser205, and Glu222; and the five water molecules closest to the chromophore. One water molecule is protonated when investigating the stability of a hydronium.

Our simulations at room temperature indicate that a hydronium is not stable. In particular, in the presence of a protonated Glu222, the additional proton migrates to the nitrogen of the imidazolinone ring, leading to the formation of a zwitterionic chromophore and a neutral water. If we hypothesize that the

Table 1. CASPT2 Vertical Excitation Energies (eV) of the Minimal Neutral Chromophore Computed with the ANO-S-PVDZ Basis on the Ground-State DFT/BLYP Geometry^a

CAS(<i>n,m</i>)	<i>N</i>	state	osc. str.	MS-PT2	SS-PT2	CASSCF	$ \Delta\mu $
2,2	2	2	0.96	3.56	3.55	4.79	2.8
4,4	2	2	0.94	3.82	3.79	4.72	4.2
6,6 ^b	3	3	0.75	3.87	4.24	5.37	5.3
8,8 ^b	3	3	0.80	3.99	4.24	5.46	5.5
10,10 ^b	3	3	0.86	3.99	4.06	5.44	5.3
12,11 ^b	3	3	0.68	3.69	3.87	5.24	3.3
	4	3, 4	0.72, 0.23	3.38	3.80	5.24	3.1
14,13 ^b	4	4	0.82	3.61	3.73	5.35	3.6
	5	4, 5	0.80, 0.16	3.49	3.75	5.32	3.6
16,14 ^b	4	4	0.89	3.60	3.64	5.31	4.6
	5	4, 5	0.85, 0.15	3.49	3.67	5.27	4.5
IPEA0							
16,14 ^b	4	4	0.89	3.17	3.23	5.31	4.6

^a Different CAS expansions of *n* electrons in *m* active π orbitals are used, and the total number of π electrons in the reference is 16. The SA calculations are on the lowest *N* roots, and the brighter states with the corresponding CASSCF oscillator strength (osc. str.) are listed. The standard IPEA shift is used in the multistate (MS) and single-state (SS) CASPT2 calculations, and the CASSCF and single-state CASPT2 excitations are relative to the state with the highest oscillator strength. We also report the CASSCF modulus of the dipole difference ($|\Delta\mu|$) in Debye. In the last line, we also list the CASPT2 excitations for the CAS(16,14) expansion computed without IPEA shift (IPEA0). ^b Imaginary shift of 0.1 au.

proton on the hydronium is originating from Glu222 and repeat the simulation with an anionic Glu222, the proton initially forming the hydronium migrates back to Glu222 during the dynamics, yielding a neutral Glu222 and a water molecule.

Finally, we find that the zwitterionic form is only stable in the presence of a protonated Glu222, while, in a doubly protonated GFP chromophore (on the phenol and imidazolinone rings) with a deprotonated Glu222, the proton jumps during the dynamics from the nitrogen of the imidazolinone to Glu222, neutralizing the carboxyl group. Therefore, the proton shuffling between a zwitterionic and a cationic form is not possible, while the existence of a zwitterionic-neutral equilibrium can be postulated only in the presence of a neutral Glu222 in both forms of the chromophore.

3. VERTICAL EXCITATION ENERGIES

The vertical excitation energies of the A and B forms of GFP are computed within CASPT2/MM, TDDFT/MM, and QMC/MM on the ground-state DFT/PBE geometry equilibrated within CPMD/Amber03. The polarization effect of the protein environment on the excitation is accounted for by point charges at the positions obtained in the QM/MM equilibration.

3.1. CASPT2 Results. We compute here the multistate (MS) and single-state (SS) CASPT2 vertical excitation energies of the neutral A and anionic B forms using the ANO-S-PVDZ basis. For comparison, we also calculate the excitation energies of the anionic and neutral minimal (*p*-HBI) and methyl-terminated (*p*-HBID) models in the gas phase using the same basis and the DFT/BLYP geometries obtained in ref 35. All excitations are computed using the standard IPEA Hamiltonian, and for the

Table 2. CASPT2 Vertical Excitation Energies (eV) of the Methyl-Terminated Neutral Chromophore Computed with the ANO-S-PVDZ Basis on the Ground-State DFT/BLYP Geometry^a

CAS(<i>n,m</i>)	<i>N</i>	state	osc. str.	MS-PT2	SS-PT2	CASSCF	$ \Delta\mu $
2,2	2	2	0.94	3.46	3.46	4.82	2.2
4,4	2	2	0.95	3.72	3.70	4.78	3.4
6,6 ^b	2	2	0.93	3.84	3.82	5.08	4.8
8,8 ^b	3	3	0.83	4.12	4.18	5.54	4.9
10,10 ^b	3	3	0.87	4.01	3.99	5.50	4.4
12,11 ^b	4	3, 4	0.41, 0.26	3.06	4.13	5.18	2.5
	5	3, 4	0.41, 0.25	2.92	4.15	5.20	2.8
14,13 ^b	4	4	0.57	3.34	3.92	5.24	1.1
	5	4, 5	0.57, 0.34	3.14	3.92	5.23	1.2
16,14 ^b	4	4	0.64	3.33	3.82	5.21	1.7
	5	4, 5	0.63, 0.33	3.15	3.85	5.20	2.0
IPEA0							
16,14 ^b	4	4	0.64	2.75	3.35	5.21	1.7

^a The total number of π electrons on the chromophore is 16. For the notation, see the caption of Table 1. ^b Imaginary shift of 0.1 au.

CAS(16,14) expansion, we also report the excitations computed without the IPEA shift to allow for a comparison with previous CASPT2 calculations.

The computation of the vertical excitations of the neutral models in the gas phase and in the protein is considerably more problematic than the CASPT2 calculations for the anionic counterparts. In Table 1, we begin with the minimal neutral model, where we can observe signs of the complications that will become evident in the neutral methyl-terminated model. In particular, the brightest state is no longer the second one at the CASSCF level, and whenever the oscillator strength is non-negligible on more than one state, we observe a marked difference between the single- and multistate CASPT2 excitations. In fact, when increasing the expansion from CAS(10,10) to CAS(12,11), the oscillator strength decreases on the bright state and becomes non-negligible on an additional state while the multi- and single-state excitations differ by as much as 0.4 eV. In CAS(12,11), we cannot stabilize the lone-pair orbital of the phenolic oxygen but must include the nitrogen lone pair on the imidazolinone while omitting the bonding and antibonding π orbitals on the benzene. Further increasing the expansion to correlate all 16 π electrons in the reference in a minimal active space appears to cure the problem. The oscillator strength in the bright state increases, and the difference between the multi- and single-state CASPT2 excitations is only 0.2 eV when the additional state with non-negligible oscillator strength is included in the multistate calculation. Overall, we note that the single-state CASPT2 excitation is well behaved and displays a smooth convergence as a function of the size of the expansion. On the contrary, the multistate excitation is particularly sensitive to the inclusion of states other than the bright one, which leads to a significant divergence of the single- and multistate values for small CAS expansions.

The CASPT2 excitations of the methyl-terminated chromophore are listed in Table 2. Since the minimal and methyl-terminated models only differ in the termination, one would expect rather similar excitations in the two cases. This expectation appears fulfilled when inspecting the single-state CASPT2 excitations, which display a similar convergence as a function of

Table 3. CASPT2/Amber99 Vertical Excitation Energies (eV) of the A Form Computed with the ANO-S-PVDZ Basis^a

CAS(<i>n,m</i>)	<i>N</i>	state	osc. str.	MS-PT2	SS-PT2	CASSCF	$ \Delta\mu $
2,2	2	2	1.11	3.29	3.26	4.50	3.7
4,4	2	2	0.89	3.54	3.49	4.36	5.1
6,6	2	2	0.83	3.72	3.69	4.59	8.8
8,8	2	2	0.83	3.72	3.69	4.60	8.9
10,10 ^b	2	2	1.03	3.56	3.53	4.52	7.5
12,11 ^b	2	2	1.10	3.56	3.56	4.21	8.3
14,13 ^b	2	2	0.99	3.57	3.56	4.16	9.8
16,14 ^b	2	2	1.00	3.53	3.53	4.16	9.9
IPEAO							
16,14 ^b	2	2	1.00	3.14	3.13	4.16	9.9

^a The ground-state DFT/PBE geometry equilibrated within CPMD/Amber03 is used. For the notation, see the caption of Table 1.

^b Imaginary shift of 0.1 au.

the CAS size and are only blue-shifted by 0.2 eV with respect to the CAS(16,14) values of the minimal model. On the other hand, the multistate excitations differ from the single-state values even more dramatically than for the minimal model. The CAS(12,11) expansion is over a set of orbitals with the same character as in the minimal models and marks the appearance of a second state with non-negligible oscillator strength while the oscillator strength of the bright state drops to half the value of the CAS(10,10) state. The difference between multi- and single-state CASPT2 excitations is now more than 1 eV. Further increasing the size of the CAS expansion ameliorates the situation, but correlating all π electrons on the chromophore in a minimal CAS(16,14) is not sufficient to reduce the difference between single- and multistate CASPT2 results, which remains larger than 0.6 eV. Since the presence of the methyl groups might be responsible for iper-conjugation effects, we also correlate the electrons of the methyl carbons by performing CAS(18,15) and CAS(20,16) calculations over four and five states (data not shown). However, we find that iper-conjugation plays no significant role, as the excitations of these larger active spaces are identical to the CAS(16,14) results.

We believe that the large difference between multi- and single-state results is due to the inability of the minimal CAS(16,14) expansion to provide a satisfactory description of all states included in the multistate calculation. While increasing the size of the expansion to a minimal CAS(16,14) was sufficient to adjust the multistate value in the minimal model, this active space does not appear to be sufficient in the presence of methyl-termination. Unfortunately, due to the high computational cost, we were not able to add a sufficiently high number of virtual orbitals in the CAS(16,14) expansion to further stabilize the multistate results. Since the chemical behavior of the methyl-terminated model should be similar to the minimal one where the multistate excitation approaches the single-state value for larger expansions, we consider the single-state CASPT2 excitations to be more reliable estimates, and we will take them as our reference data. Moreover, it has been previously observed that large deviations of the multistate CASPT2 excitations from their single-state counterparts should be taken with caution and most likely are an indication of the failure of the multistate approach.^{43,44}

The CASPT2 excitations of the A form in the presence of the MM charges are collected in Table 3. Differently from the gas-phase models, it is always possible to obtain a bright state with a

Table 4. CASPT2 Vertical Excitation Energies (eV) of the Minimal Anionic Chromophore Computed with the ANO-S-PVDZ Basis on the Ground-State DFT/BLYP Geometry^a

CAS(<i>n,m</i>)	osc. str.	MS-PT2	SS-PT2	CASSCF	$ \Delta\mu $
2,2	1.29	2.68	2.67	3.99	0.3
4,4	1.36	2.96	2.96	3.45	1.4
6,6	1.19	2.93	2.90	3.30	2.7
8,8	1.15	2.96	2.90	3.33	3.4
10,10	1.17	2.96	2.85	3.26	4.2
12,11	1.39	2.82	2.82	3.42	1.5
14,13	1.35	2.80	2.79	3.16	1.0
16,14	1.34	2.78	2.77	3.13	0.9
IPEAO					
16,14	1.35	2.57	2.56	3.13	0.9

^a All SA calculations are on the two lowest roots. For the notation, see the caption of Table 1.

Table 5. CASPT2 Vertical Excitation Energies (eV) of the Methyl-Terminated Anionic Chromophore Computed with the ANO-S-PVDZ Basis on the Ground-State DFT/BLYP Geometry^a

CAS(<i>n,m</i>)	osc. str.	MS-PT2	SS-PT2	CASSCF	$ \Delta\mu $
2,2	1.30	2.65	2.64	4.01	0.1
4,4	1.37	2.92	2.92	3.48	1.7
6,6	1.19	2.92	2.88	3.33	3.4
8,8	1.15	2.95	2.89	3.37	3.8
10,10	1.18	2.94	2.83	3.31	4.6
12,11	1.18	2.91	2.82	3.28	4.4
14,13 ^b	1.35	2.79	2.78	3.18	1.4
16,14 ^b	1.35	2.77	2.76	3.15	1.7
IPEAO					
16,14 ^b	1.35	2.34	2.33	3.15	1.7

^a All SA calculations are on the two lowest roots. For the notation, see the caption of Table 1. ^b Imaginary shift of 0.1 au.

large oscillator strength and with excitations that are similar in the single- and multistate calculations even though the orbitals in the various CAS expansions closely resemble the ones in vacuo. We note however that also for the A form we find energetically close CASSCF minima characterized by significantly different multi- and single-state excitations and oscillator strength spread over many states. For instance, even for the largest CAS(16,14) expansion, one can obtain a minimum where the CASSCF excitation is higher by 0.5 eV and the multistate CASPT2 value as low as 3.12 eV.

In Tables 4, 5, and 6, we report the CASPT2 excitations of the anionic minimal, methyl-terminated, and B-form models, respectively. The behavior of the excitations is rather similar for the three anionic models, which are characterized by a bright state being clearly dominant and a close agreement between the single- and multistate results. For all anionic models, the CAS(12,11) expansion is constructed by omitting the lone-pair orbital on the imidazolinone nitrogen and the π bonding and antibonding orbitals on benzene.

For all models and the chemically relevant CAS(16,14) expansion correlating all π electrons on the chromophore, we

Table 6. CASPT2/Amber99 Vertical Excitation Energies (eV) of the B Form Computed with the ANO-S-PVDZ Basis^a

CAS(<i>n,m</i>)	osc. str.	MS-PT2	SS-PT2	CASSCF	$ \Delta\mu $
2,2	1.30	2.73	2.69	4.06	1.6
4,4	1.34	2.96	2.96	3.57	4.4
6,6	1.35	2.98	2.98	3.69	4.6
8,8	1.21	3.02	2.95	3.60	6.4
10,10	1.17	2.91	2.87	3.45	6.4
12,11	1.36	2.84	2.84	3.52	5.0
14,13 ^b	1.28	2.84	2.84	3.26	5.9
16,14 ^b	1.30	2.82	2.82	3.20	5.8
IPEAO					
16,14 ^b	1.30	2.38	2.38	3.20	5.8

^aAll SA calculations are on the two lowest roots. The ground-state DFT/PBE geometry equilibrated within CPMD/Amber03 is used. For the notation, see the caption of Table 1. ^bImaginary shift of 0.1 au.

compute the CASPT2 excitations also without the IPEA shift and therefore with the old definition of zero-order Hamiltonian. The resulting excitations are significantly red-shifted by as much as 0.4–0.5 eV for all models. The only system where the deviation is smaller and equal to 0.2 eV is the minimal anionic model, which indicates that the minimal CAS on all π electrons captures the most relevant correlation effects for this model.

Finally, we also list the modulus of the difference between the CASSCF dipole moments of the ground and the bright excited state for all models and CAS expansions. The magnitude of the dipole moment difference is closely related to the charge transfer character of the excitation, which we estimate as the change in the Mulliken charges induced by the excitation (see Table 1 of Supporting Information). For the anions, the modulus of the dipole difference of the CAS(16,14) states increases from 0.9 to 1.7 to 5.8 D when going from the minimal to the methyl-terminated model to the B form. Correspondingly, the degree of charge transfer from the phenolic ring to the carbon bridge increases from the minimal to the methyl model, and the transfer becomes then predominantly to the imidazolinone ring in the B form. For the neutral moieties, the dipole difference decreases from 4.6 to 1.7 D from the minimal to the methyl model and increases to as much as 9.9 D in the A form. Differently from the gas-phase models, the charge transfer of the A form is from the phenolic ring mainly toward the central bridge. We note that, in all cases, the CASSCF magnitude of the dipole moment difference is very close to the value computed within CASPT2.

3.2. TDDFT Results. We compute the TDDFT excitations of the anionic and neutral methyl-terminated models and of the A and B forms using the aug-cc-pVDZ basis and several exchange-correlation functionals, namely, BLYP, B3LYP, CAM-B3LYP, LC-BLYP, and LC- ω PBE. As shown in Table 2 of the Supporting Information, the inclusion of augmentation yields a significantly faster basis-set convergence for the excitations and dipole moments, and the use of the aug-cc-pVDZ gives results converged within 0.01 eV for both the anionic and neutral states.

We collect the excitations of the neutral and anionic methyl-terminated models in Table 7. For the anion, all functionals yield excitation energies in the range 3.0–3.1 eV with the exception of the generalized gradient approximation BLYP, which gives a red-shifted value of 2.8 eV. For the neutral methyl-terminated model, the spread of results is larger, with BLYP giving a very low

Table 7. TDDFT Vertical Excitation Energies (eV) of the Methyl-Terminated Neutral and Anionic Chromophore Models Computed with Different Functionals and the aug-cc-pVDZ Basis on the Ground-State DFT/BLYP Geometry^a

functional	neutral		anion	
	E_{exc}	$ \Delta\mu $	E_{exc}	$ \Delta\mu $
BLYP	3.06 (0.51)	1.3	2.79 (0.73)	2.8
B3LYP	3.33 (0.64)	0.4	2.96 (0.90)	1.7
CAM-B3LYP	3.56 (0.71)	0.6	3.05 (1.02)	1.2
LC-BLYP	3.79 (0.78)	1.0	3.10 (1.11)	1.2
LC- ω PBE	3.74 (0.75)		3.08 (1.09)	

^aWe also report the oscillator strength in brackets and the modulus of the dipole difference ($|\Delta\mu|$) in Debye.

Table 8. TDDFT/Amber99 Vertical Excitation Energies (eV) of the Neutral A and Anionic B Forms Computed with Different Functionals and the aug-cc-pVDZ Basis^a

functional	A form		B form	
	E_{exc}	$ \Delta\mu $	E_{exc}	$ \Delta\mu $
BLYP	3.07 (0.47)	3.8	2.82 (0.79)	2.6
B3LYP	3.22 (0.79)	1.5	3.00 (0.97)	1.5
CAM-B3LYP	3.42 (0.86)	2.5	3.10 (1.05)	2.0
LC-BLYP	3.61 (0.91)	3.3	3.17 (1.12)	2.8
LC- ω PBE	3.57 (0.90)		3.15 (1.11)	

^aWe also report the oscillator strength in brackets and the modulus of the dipole difference ($|\Delta\mu|$) in Debye. The ground-state DFT/PBE geometry equilibrated within CPMD/Amber03 is used.

excitation energy and LC-BLYP the highest one. The use of long-range corrected functionals yields excitation energies in the range of 3.6–3.8 eV characterized by a larger oscillator strength and by dipole moments rather similar to the anionic values.

As shown in Table 8, the TDDFT excitations in the protein for the neutral A and anionic B forms follow a pattern similar to the one in the gas-phase counterparts. The BLYP excitations are red-shifted with respect to the values obtained with the other functionals, and the difference is more marked for the neutral case where the oscillator strength within BLYP is smaller and spurious excitations appear at lower energies (not shown in the table). Long-range corrected functionals yield excitations in the range of 3.4–3.6 eV and of 3.1–3.2 eV for the neutral and anionic forms, respectively. Again, the difference between the dipole moments of the neutral and anionic forms is not very large when employing hybrid or long-range corrected functionals. The change in the dipole moment induced by excitation is however generally larger in the protein environment than in the gas-phase for both neutral and anionic moieties.

3.3. QMC Results. We compute the variational (VMC) and diffusion Monte Carlo (DMC) excitations of the A and B forms in the presence of the MM environment, using the D+ basis set. In the Jastrow–Slater wave functions, we employ CASSCF expansions expressed over natural orbitals and truncated with appropriate thresholds on the CSF coefficients. Since the optimization of the orbitals does not significantly affect the QMC excitations of the anionic models in the gas phase,³⁵ we only optimize here the Jastrow and linear coefficients in energy

Table 9. QMC/Amber03 Vertical Excitation Energies (eV) of the Neutral A and Anionic B Forms Computed with the D+ Basis^a

	CAS(<i>n,m</i>)	Thr.	Det/CSF	VMC	DMC
B form	2,2	0.00	4/3	3.35(7)	3.02(8)
	12,11	0.10	7/4	3.25(7)	3.1(1)
		0.05	80/28	3.4(1)	3.1(1)
A form	2,2	0.00	4/3	3.99(7)	3.78(9)

^aThe statistical error is indicated in brackets. We use a CAS(2,2) and a CAS(12,11) expansion in the determinantal component truncated with different thresholds (Thr.) on the CSF coefficients and report the number of determinants and CSFs retained in the expansion. The ground-state DFT/PBE geometry equilibrated within CPMD/Amber03 is used.

minimization within variational Monte Carlo in a state-averaged manner.

In Table 9, we report the VMC and DMC excitations of the B form computed with a CAS(2,2) and a larger CAS(12,11) expansion shown to be adequate in the convergence of the CASPT2 excitation. We find that the excitations computed with different CAS expansions and thresholds are equivalent within statistical error at both the VMC and the DMC levels, and even a simple CAS(2,2) wave function appears to be sufficient. The more reliable DMC excitation is systematically lower than the VMC value and close to the DMC estimate of 3.04(4) eV for the methyl-terminated model in the gas phase.³⁵ For the A form, we only compute the QMC excitation with a CAS(2,2) calculation within GAMESS to the same CASSCF minimum reported in Table 3. Nevertheless, the DMC excitation obtained with a CAS(2,2) is an indication that the DMC estimate is similar to CASPT2 and TDDFT with long-range corrected functionals.

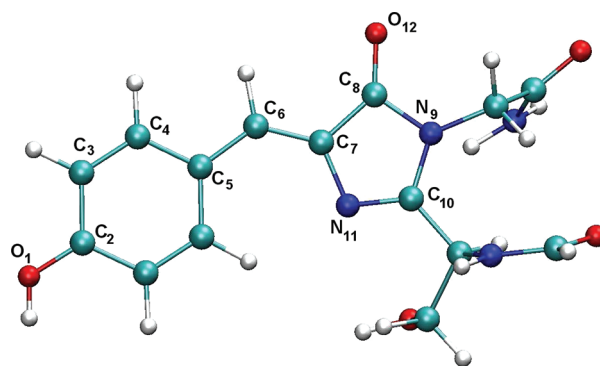
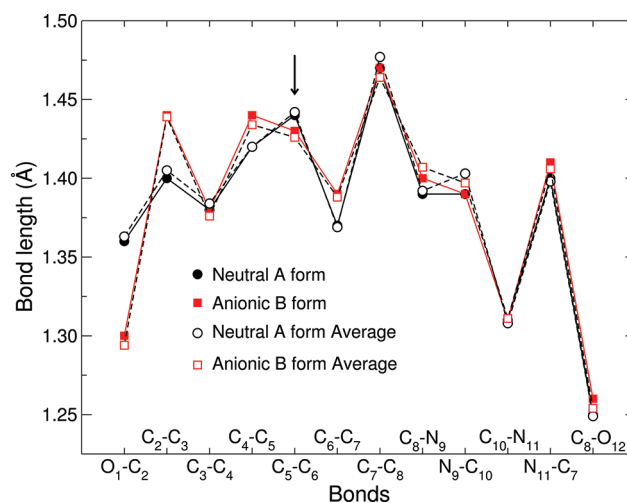
4. DISCUSSION

In this section, we first discuss the features of our structural models and compare them with other simulations available in the literature. We describe the tests we performed to investigate the reliability of our structures with respect to the choice of QM theoretical approach and MM force field. We also discuss our findings on the stabilization of alternative nonstandard protonations and of a solvated hydronium in proximity of the chromophore.

We then focus on the relative performance of the theoretical approaches employed to compute the vertical excitation energies of the neutral A and anionic B forms and then discuss their comparison with the experimental absorption spectra.

4.1. Structural Analysis of GFP Models. We discuss here the structural features of the chromophore for the neutral and anionic forms of wild-type GFP obtained in our QM/MM calculations. For the labeling of the atoms of the chromophore, we refer the reader to Figure 3.

The structure of the chromophore is expected to play a very important role in tuning its excited state properties. In particular, the degree of bond-length alternation in the conjugate chain running through the chromophore will be correlated to the size of the excitation, which is here of π to π^* character. In addition, the local features of the binding site of the chromophore can affect the spectral response of the chromophore by either tuning the internal geometrical structure of the chromophore or as a polarizing environment.

**Figure 3.** Atom numbering used for the chromophore of GFP. This image is produced with VMD.**Figure 4.** Bond lengths of the chromophore of the neutral A and anionic B forms as obtained in our PBE/Amber CPMD simulations. The bond lengths of the chromophore resulting from our annealing procedure are compared with the average values obtained along a room-temperature QM/MM molecular dynamics simulation. The root-mean-square fluctuations on the bond lengths are on the order of 0.02–0.03 Å, and the error bars on the averages are smaller than the size of the symbols. The C₅–C₆ bond in the central carbon bridge is indicated with an arrow.

Before analyzing the structural differences between the various forms, it is important to verify that the models obtained in our annealing procedure are indeed representative of an average configuration of the chromophore within the protein. To this end, we compare our annealed models of the A and B forms to the bond lengths obtained by averaging over approximately 1.5 ps of a molecular dynamics QM/MM simulation at room temperature. As shown in Figure 4, the annealed and average bond lengths are remarkably similar for both forms, with differences smaller than 0.01 Å, demonstrating the validity of employing our annealed structures as representative models in further analysis and in the calculations of the excitation properties of both forms.

To understand the geometrical effects of the protein environment on the chromophore, the bond lengths of the chromophore in the neutral and anionic forms are compared with the values optimized in vacuo in Figure 5. In the protein, we observe a slight shortening of about 0.01 Å in the bond lengths of the neutral A form, with larger deviations in the imidazolinone ring close to the

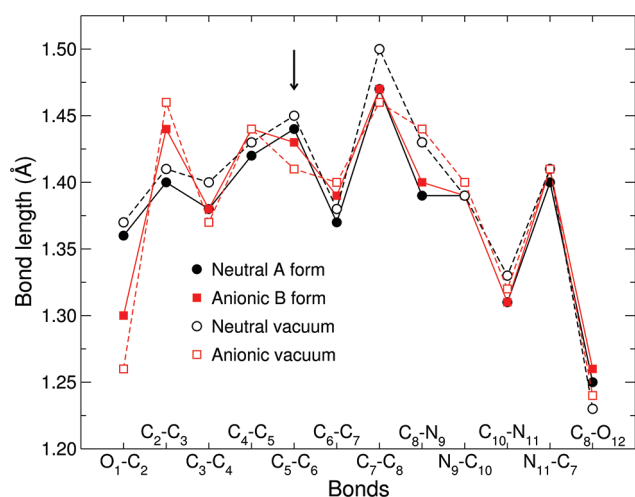


Figure 5. Bond lengths of the chromophore of the neutral A and anionic B forms as obtained in our PBE/Amber CPMD simulations. The results for the chromophore models optimized in vacuo with BLYP/cc-pVTZ are also shown. The C_5-C_6 bond in the central carbon bridge is indicated with an arrow.

Arg96 counterion. Similarly, the bond lengths of the anionic B form do not dramatically differ from the values we obtain in vacuo. The most significant difference is observed in the degree of bond alternation, which increases close to the central bridge of the anionic form. Moreover, the O_1-C_2 bond is lengthened in the protein due to hydrogen bonding to close-by residues.

When comparing the neutral and the anionic forms, we see that, in vacuo, the largest difference occurs in proximity of the O_1-C_2 bond. In fact, as expected, after deprotonation, it loses its single-bond character and significantly shortens by about 0.1 Å. As a consequence, the aromaticity (similar bond lengths) of the phenolic ring is reduced in the anionic form with respect to the neutral case. Finally, the degree of bond alternation in proximity of the central bond is smaller in the anion than in the neutral chromophore. Interestingly, we note that the difference between the neutral and the anionic bond lengths is overall smaller in the protein. In particular, the degree of bond alternation close to the bridge is more similar between anion and neutral form in the presence of the protein. Overall, it appears that the protein environment acts to partially compensate the change in protonation state and keeps the chromophore in a more similar structural conformation in the two protonation states with respect to a vacuum.

In Figure 6, we compare the geometrical properties of the chromophore of the three forms of GFP with the results of other simulations available in the literature. We first focus on the AM1/CHARMM calculations by Marques et al.,⁴⁵ as the structural features of their models most significantly differ from ours as well as from other calculations reported in the literature. We first note that Marques et al. construct the I form by deprotonation of the neutral form although they refer incorrectly to this structure as the B form. Their neutral A and anionic I forms are rather similar with the exception of the O_1-C_2 bond, which is significantly shorter in the I form, in agreement with our calculations. Both structures display however a more marked bond-length alternation along the chromophore than our models. This is particularly evident for the I form where the bond alternation in the central carbon bridge is 0.09 Å compared to 0.05 Å in our calculations.

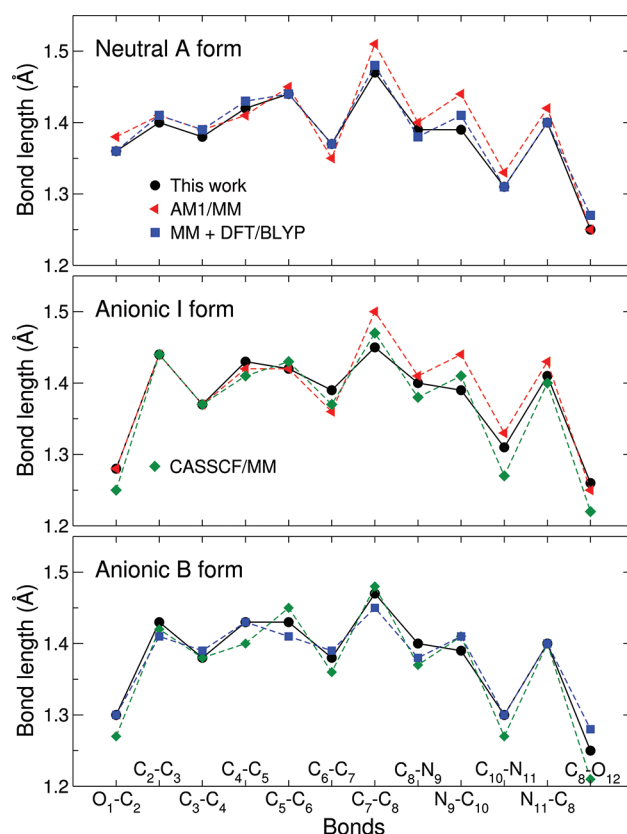


Figure 6. Bond lengths (Å) of the chromophore of the A, I, and B forms of wild-type GFP obtained with PBE/Amber in CPMD. We also show the results of the AM1/MM,⁴⁵ CASSCF/MM,⁴⁶ and MM simulations followed by partial QM relaxation⁴⁷ (denoted as MM+DFT/BLYP). The bonds of the central bridge are C_5-C_6 and C_6-C_7 .

Another clear difference is the bond length of the subsequent single carbon bond of the imidazolinone ring, C_7-C_8 , which is 0.05 Å longer than our value for both the neutral and anionic I forms.

Our structures for the neutral A and anionic B forms are instead in close agreement with the simulations of Laino et al.⁴⁷ denoted here as “MM + DFT/BLYP”. The authors employ the Amber98 force field and parametrize the chromophore accordingly, to perform a first classical MD relaxation of the protein. They then refine the structural parameters with a DFT/BLYP optimization of an isolated cluster comprising the chromophore and its close residues with fixed heavy atoms at the boundary.

Finally, we compare our anionic I and B forms with the results of the CASSCF/CHARMM calculations by Sinicropi et al.,⁴⁶ who relax the chromophore structure within CASSCF together with three closeby classical waters. The rest of the protein is kept to the original crystallographic coordinates, and the I form is obtained starting from the A form and manually reorienting the residues Ser205 and Glu222 to form the expected hydrogen bond network. The B form is derived from the I form by rotating and relaxing Thr203 to create a hydrogen bond with the phenolic oxygen of the chromophore. Despite the lack of complete relaxation, their hydrogen-bond distances are in reasonable agreement with our values. As generally found when comparing CASSCF and DFT structures, their chromophore displays a somewhat larger degree of bond-length alternation as compared to our DFT calculations.

To understand the origin of some of the differences with previous studies, we investigate how particular choices in the description of

the QM and MM parts affect the structural features of the chromophore. We first consider the effect of the choice of the force field since the AM1⁴⁵ and CASSCF⁴⁶ calculations use CHARMM while we employ the Amber03 parametrization. Starting from our PBE/Amber03 annealed model of the B form, we optimize the chromophore alone with BLYP/CHARMM for fixed positions of the MM charges. We note here that BLYP and PBE lead to equivalent structures in the gas phase and the use of BLYP/CHARMM instead of PBE/CHARMM is simply due to technical reasons.⁴⁸ As shown in Table 3 of the Supporting Information, we find that the chromophores optimized with BLYP/CHARMM and PBE/Amber03 are equivalent with bond lengths deviating at most by 0.02 Å. Therefore, the choice of force field does not lead to significant differences in the main structural features and cannot explain the large deviations observed between the AM1/CHARMM model of ref 45 and our structure.

We then explore whether the use of the CASSCF method in optimizing the QM part and/or the lack of full relaxation in their protein model are responsible for the differences between the CASSCF/CHARMM structure of ref 46 and our model of the B form. Starting from the coordinates of ref 46, we optimize the chromophore within DFT/BLYP at fixed MM positions and, as shown in Table 3 of the Supporting Information, recover a chromophore which closely resembles the structure obtained in our very different DFT-based annealing procedure. Therefore, the observed differences must be ascribed to the use of the CASSCF approach to optimize the QM part and not to the details of the construction of the protein model. Then, we investigate how the choice of the exchange-correlation functional affects the final QM geometry. In particular, we consider the use of hybrid functionals and, starting from our model of the B form, we perform a B3LYP/Amber99 optimization of the chromophore at fixed positions of the MM charges. Even in this case, the resulting chromophore has similar bond lengths to the PBE/Amber03 values. We already note here and discuss further in the next section that the observed geometrical differences between the CASSCF, DFT/B3LYP, and our DFT/BLYP structures do not result in substantial differences in the TDDFT/MM excitations of the B form.

To further probe the reliability of our structures, we perform few steps of geometrical optimization using the CASPT2/MM method starting from our model of the B form. We employ the Amber99 force field, a CAS(12,11) expansion in the CASSCF wave function, and keep the MM charges at the fixed positions of our CPMD/MM annealed structures. We converge the CASPT2 root-mean-square fluctuations of the Cartesian coordinates and forces to less than 0.01 Å and 0.01 au, respectively, and the chromophore remains very similar to the DFT starting structure with deviations smaller than 0.01 Å. Therefore, this test gives us confidence in our choice of DFT/BLYP being able to capture the correct structural properties of the chromophore embedded in the protein.

Finally, we perform extensive QM/MM simulations of the anionic I form increasing the QM part to include residues surrounding the chromophore. The structure is first equilibrated at room temperature with CPMD/Amber03 within DFT/PBE, and the QM part consists of Arg96, Gln94, and the chromophore. The QM component is then enlarged to also include Glu222, His148, and the water close to the phenolic oxygen, and the system is slowly annealed to zero temperature. The resulting chromophore deviates by less than 0.01 Å from the original geometry obtained with only the chromophore in the QM part. These

Table 10. Vertical Excitation Energies (eV) of the Neutral and Anionic Methyl-Terminated Models in the Gas Phase, and the Neutral A and Anionic B Forms in the Protein^a

	Gas phase		Protein	
	Neutral	Anion	Neutral	Anion
CAM-B3LYP	3.56	3.05	3.42	3.10
LC-BLYP	3.79	3.10	3.61	3.17
CASPT2	3.82	2.76	3.53	2.82
DMC	—	3.04(4) ^b	—	3.1(1)
Expt.	3.51	2.59 ^c , 2.75 ^d , >2.60 ^e , 2.84 ^f	3.05 ^g	2.63 ^g

^aThe TDDFT excitations are reported for the CAM-B3LYP and LC-BLYP functionals and the aug-cc-pVDZ basis set. The best available single-state CASPT2 and DMC excitations, obtained with the ANO-S-PVDZ and D+ bases, respectively, are shown. The ground-state DFT/BPE geometry equilibrated within CPMD/Amber03 is used. ^bDMC excitation from ref 35. ^cPhotodestruction experiments of ref 52. ^dPhotodestruction experiments of ref 53. ^ePhotodestruction experiments of ref 54. ^fExtrapolation of Kamlet–Taft fit of solution experiments.⁵ ^gAbsorption maxima in the protein at 1.7 K.⁴

simulations are also repeated including dispersion-corrected atom-centered (DCACP) pseudopotentials.^{49,50} The use of DCACP does not lead to significant changes in the structure.

In summary, our structures constructed within CPMD/MM via a room-temperature MD procedure followed by annealing appear to be reliable. They are representative of an average configuration of the chromophore–protein complex as demonstrated by a comparison with the average bond lengths over room-temperature MD trajectories. Treating only the chromophore in the QM part is sufficient, and the geometry is not sensitive to the particular choice of nonpolarizable force field in the MM part, namely, CHARMM or Amber. The choice of DFT to treat the QM component is appropriate, as a partial optimization within CASPT2/MM approach leaves the structure unchanged. We can rationalize the small difference with the CASSCF structures of ref 46 as due to the use of a different QM method in the optimization. Finally, our structures compare well to the ones obtained in the MM simulations refined by further QM relaxation of ref 47. The difference with the AM1/MM structure of ref 45 remains the most significant and is possibly due to the positive protonation state of most His residues in their model.⁵¹

4.2. Comparison of Theoretical and Experimental Excitations. In Table 10, we summarize the most representative theoretical results for the vertical excitations of the neutral and anionic methyl-terminated models in the gas phase and of the neutral A and anionic B forms in the protein. We report the single-state CASPT2 excitations computed with the largest CAS(16,14) expansion since the single-state values are either close to their multistate counterpart or appear to be a more reliable estimate of the excitation energy as shown in the case of the neutral gas-phase models. We list the TDDFT excitations computed with the CAM-B3LYP and LC-BLYP functionals, as calculations with BLYP are in some cases plagued by the appearance of spurious low-lying excitations. Finally, we consider the best DMC values available. Our theoretical estimates are compared with the available experimental results, namely, three different photodestruction experiments for the anion in the gas phase,^{52–54} the extrapolation to vacuum conditions of solution experiments for the neutral and anionic species,⁵ and the low-temperature absorption spectra in the protein.⁴

We first focus on the results for the anionic models, which are remarkably consistent among the wave function and DFT methods both in the gas phase and in the protein. In the gas phase, the excitations are between 2.76 and 3.10 eV, in agreement with our previous calculations,³⁵ which are here extended to include a more complete set of DFT functionals. The CASPT2 excitations are computed with the ANO-S-VDZP basis for consistency with the calculations in the protein, so small differences with the CASPT2 estimates of ref 35 must be attributed to the use of a different basis set. We also list the DMC excitation in the gas phase computed without augmentation and a CAS(8,8) expansion and recall here that tests on the inclusion of augmentation and larger CAS spaces indicate that the DMC excitation of the anionic model is rather robust and consistently at about 3 eV.³⁵

The computation of the anionic excitation in the gas phase is in principle complicated by the fact that the excitation lies above the ionization threshold, as demonstrated in recent photodestruction experiments.⁵³ Nevertheless, the DFT value is rather insensitive to the use of very large basis sets,⁵⁵ and even artificially raising the ionization threshold well above the excitation via asymptotically corrected potentials (e.g., the statistically average orbital potential approach) does not affect its value.³⁵ Similarly, the DMC estimate appears to be largely unaffected by the choice of basis and CAS expansions.³⁵ These findings indicate that this metastable state in the continuum is dominated by a well localized π to π^* transition, which renders possible the computation of this excitation in the continuum. Experimentally, early photodestruction experiments assign the vertical transition to the only observed peak in the absorption spectrum at 2.59 eV.⁵² A second photodestruction experiment resolves however a rather different shape of the absorption spectrum and reconsiders its interpretation assigning the adiabatic transition to the lowest peak at 2.59 eV and the vertical transition to the second, newly resolved peak at 2.75 eV.⁵³ The most recent photodestruction experiment reveals a strong dependence of the shape of the spectrum on the excitation laser power and leads the authors to conclude that such a spectrum cannot reliably represent the optical absorption spectrum of the anionic GFP model in the gas phase.⁵⁴ Finally, the multivariant Kamlet–Taft fit of the absorption maxima in solution in terms of the acidic, basic, and polar solvating parameters extrapolates to 2.84 eV for the conditions of a vacuum.⁵ In view of the apparent uncertainty on the shape of the spectra in photodestruction experiments, our theoretical estimates appear to be in the right range as they fall between 2.76 and 3.10 eV and display a blue shift which can be easily explained with vibronic effects and with the intrinsic theoretical limitations of predicting a metastable excitation in the continuum.

In the protein, the theoretical excitations of the anionic B form range between 2.82 and 3.17 eV and therefore display a negligible bathochromic shift with respect to the gas phase. Our theoretical estimates appear to be rather robust and even more insensitive to the internal parameters of the theory than the calculations in the gas phase. In particular, the excited state is not strongly multiconfigurational as the CASPT2 excitations converge rapidly with the size of the expansion and, for the DMC values, even a small CAS(2,2) expansion is sufficient. The DFT estimates are clustered around 3.1 eV, and the presence of the protein raises the ionization threshold (estimated as minus the HOMO eigenvalue) well above the excitation, in agreement with the findings of ref 56. Therefore, with respect to the anion in the gas phase, we no longer have the complications of describing an excitation in

the continuum. The theoretical excitations computed within our QM/MM model are however significantly higher than the experimental absorption maximum of 2.63 eV measured at 1.7 K. Therefore, while experiments indicate that there should be a red shift on the order of 0.2 eV when going to the protein, theory obtains a shift less than 0.1 eV in the opposite direction, which brings the vertical excitation farther away from the location of the absorption maximum in the protein. To explain our findings within the protein, we cannot invoke vibronic or temperature effects, as the chromophore is tightly held in a rigid position within the protein pocket and the experimental absorption band of the B form is rather narrow, with the location of the maximum moving only to 2.59 eV when the measurement is performed at room temperature.

As already mentioned, the computation of the excitations of the neutral moieties is even more challenging. For instance, the CASPT2 calculations in the protein present difficulties due to the existence of multiple close minima at the CASSCF level which are characterized by different ordering of the states, oscillator strengths, and CASSCF excitations, even when all π electrons on the chromophore are correlated in a minimal active space. Both in the gas phase and in the protein, whenever the oscillator strengths are spread over multiple states, the single- and multi-state excitations differ significantly, and this behavior is particularly evident in the gas-phase calculations of the methyl-terminated model. Finally, the DFT excitations in the gas phase and in the protein computed using the generalized gradient approximation BLYP and hybrid B3LYP are significantly red-shifted with respect to all other values. Focusing on our best theoretical estimates of Table 10, the vertical excitations in the gas phase range between 3.56 and 3.82 eV and are in reasonably good agreement with the Kamlet–Taft extrapolation of solution experiments to vacuum conditions at 3.51 eV.⁵ Again, we need to recall that vibronic effects are expected to be very strong for the chromophore in the gas phase (and in solution) and will lead to an absorption maximum red-shifted with respect to the location of the vertical excitation. Nevertheless, both experiments and theoretical estimates in the gas phase unequivocally indicate that one must expect a significant bathochromic shift of at least 0.5 eV since the absorption maximum in the protein is located at 3.05 eV in measurements at 1.7 K. Our theoretical excitations of the A form lie instead in the range 3.42–3.53 eV, therefore significantly blue-shifted with respect to experiments, and display a too small bathochromic shift of 0.2–0.3 eV. Similarly to the B form, we do not expect vibronic or temperature effects to be of significant magnitude to explain the discrepancy between theoretical and experimental data in the protein. It therefore appears that, while theory can correctly predict the absorption in the gas phase of the neutral moiety, we cannot correctly reproduce the location of the vertical excitation in the protein and fail to see the large red shift due to the protein environment.

To explore possible origins of the discrepancy between the theoretical excitations and the experimental data, we follow two routes. On the one hand, we consider extended QM regions including the chromophore and the amino acids involved in the hydrogen-bond network surrounding the chromophore, namely, Ser205, Glu22, Arg96, Gln94, and the water close to the phenolic oxygen. On these extended models, we perform TDDFT/MM calculations with the CAM-B3LYP and LC-BLYP functionals. We find that both functionals give a red shift for the A form of about 0.15 eV. However, the corresponding excited states show non-negligible contributions of electronic transition with strong

charge transfer character also to the boundary of the QM region. Consequently, we expect that these spurious transitions will depend on the specifics of the QM/MM border and are therefore an artifact of approximate TDDFT even with the use of long-range corrected functionals. Unfortunately, it is computationally very demanding to perform such large calculations with the highly correlated approaches we consider in this paper. Therefore, as a second route, we only attempt a similar test on a QM model including only the chromophore and the counterion, Arg96, at the CASPT2/MM level and choose to explore the B form as it displays a faster and more robust convergence with CAS expansion. We find that the inclusion of a quantum counterion does not shift the excitation. In addition, attempts to stabilize a charge transfer excitation by including acceptor orbitals on the counterion in an extended CASSCF expansion are not successful as the excitation remains confined on the chromophore.

If we compare with previous calculations of the A and B form in the literature, we find the TDDFT/LDA results by Marques et al.,⁴⁵ which report an excellent agreement with the experimental absorption maxima in the protein. It is however rather difficult to assess the quality of these calculations. As already mentioned in the previous section, we need to bear in mind that they compute the I and not the B form and that the structures of their anionic and neutral chromophores are the only ones in the literature which depart by a significant extent from our models. These differences might be due to their choice of a positive protonation of most His residues in their models. Moreover, they extract the chromophores from the protein and compute the excitation without including the environment. Finally, the use of the LDA functional will naturally lead to red-shifted values as even generalized gradient approximations are affected by this problem (see section 3.2).

For the B form, previous CASPT2 calculations⁴⁶ report an excitation of 2.81 eV, blue-shifted with respect to the experimental value of 2.63 eV and in apparent agreement with our calculations. The agreement is however in part fortuitous since the authors of ref 46 employ the CASSCF method to optimize the structure, and the older definition of the zero-order Hamiltonian without IPEA shift and a smaller 6-31G* basis in their CASPT2/MM calculations. As shown in section 3.3, setting the IPEA shift to zero lowers our CASPT2 excitation of the B form from 2.82 to 2.38 eV, and the lower value of 2.38 eV should be compared to the result of ref 46. While we cannot ascribe this discrepancy to the setup of the protein model (see section 4.1), the difference between the two CASPT2 results when using the same zero-order Hamiltonian must be due to the combined use in ref 46 of a poorer basis set and the CASSCF geometry for the chromophore since both factors lead separately to a blue shift.

For the A form, Hagasawa et al.⁵⁷ report an excitation energy of 3.21 eV obtained with SAC-CI and QM/MM, in reasonable agreement with experiments. In their protein model, the QM chromophore is optimized in B3LYP/Amber, and the MM atoms are fixed at the X-ray structure. For the methyl-terminated chromophore in the gas phase, they obtain however an excitation of 3.23 eV, which is on top of the value in the protein and significantly red-shifted with respect to our estimates of 3.6–3.8 eV and to the extrapolation of solution experiments to vacuum conditions at 3.51 eV.⁵ Consequently, their SAC-CI calculations do not yield the desired bathochromic shift of at least 0.5 eV. Given the failure of their approach to reproduce the gas phase absorption and the correct bathochromic shift, the apparent agreement in the protein might be fortuitous.

Finally, in a recent study by Krylov et al.,⁵⁶ the B form is investigated using the SOS-CIS(D) method within QM/MM. The excitation energy is estimated as 2.70 eV for a protein model optimized by PBE0/Amber. The bond lengths of the chromophore are rather similar to the ones of our model, and the discrepancy between the SOS-CIS(D) and our estimates in the higher-energy range of 2.82–3.17 eV is most likely due to the use of a different technique to compute the excitation energy.

5. CONCLUSIONS

In the gas phase, DFT and wave function methods show an overall good agreement with the extrapolation of solution experiments to vacuum conditions both for the anionic and neutral species. For the anion, the vertical excitation in the gas phase ranges between 2.8 and 3.1 eV, and our theoretical estimates appear to be rather robust despite the inherent difficulty of computing an excitation in the continuum. We also note that the comparison with photodestruction experiments is difficult given the uncertainty in the shape of the spectra.^{52–54} For the neutral moiety, our calculations using either long-range corrected DFT functionals or wave function methods lead to vertical excitations in the range of 3.6–3.8 eV.

When introducing the protein environment, we do not obtain the proper bathochromic shift to the experimental absorption maxima of the anionic B and neutral A forms, which are located at 2.63 and 3.05 eV, respectively. For the anionic B form, we obtain theoretical excitations between 2.8 and 3.2 eV, therefore slightly blue-shifted with respect to the gas phase. For the neutral A form, we find excitations in the range of 3.4–3.6 eV, which are not sufficiently red-shifted. Similarly to the gas phase, wave function and DFT methods agree when long-range corrected functionals are used. Therefore, while the various theoretical approaches predict a shift between the excitations of the A and the B form (in the range 0.32–0.71 eV as compared to the experimental value of 0.4 eV), they fail rather dramatically in estimating the location of the absorption maxima.

Then, why does our model GFP not yield the desired bathochromic shift? Here, we have adopted the generally accepted protonation of the amino acids in wild-type GFP and a standard QM/MM prescription in the construction of the protein models. Within this paradigm, we can exclude the source of the problem lying in the computational details of the construction procedure. In particular, our extensive tests indicate that our structures are rather robust. Temperature effects do not play a role. The use of DFT as QM method is sufficiently accurate, and extending the QM region beyond the chromophore is not necessary when relaxing the protein. Finally, the structures are insensitive to the particular choice of nonpolarizable force field. Since the theoretical techniques to compute the excitation spectrum appear to be reliable in the gas phase and display an overall agreement among each other in the protein, we also rule out that the origin of the problem lies in our choice of QM method to compute the excitation.

Consequently, few possible sources of error require more extensive investigations. One possibility is that some amino acids in the protein are differently protonated than what is commonly accepted in the literature. Here, following the experimental suggestion of ref 42, we have investigated the stability of a hydronium in proximity to the chromophore, as the presence of this charged moiety could lead to significant structural and electronic changes in the model. However, the hydronium is not stable and

always donates the additional proton either to the chromophore or to Glu222 during a room-temperature QM/MM molecular dynamics simulation. Further investigations of other residues in the second shell surrounding the chromophore are needed. An alternative origin of the disagreement between theory and experiments is possibly the theoretical description of the polarization field of the protein on the excited state of the chromophore. While we find that the excitation is insensitive to the choice of the particular nonpolarizable force field used to describe the MM region, we cannot exclude that a more accurate description of the protein would lead to significantly different excitations. Ideally, one would extend the QM region well beyond the chromophore and the surrounding amino acids. Unfortunately, this is currently prohibitive with the highly correlated methods employed here, and our tests indicate that TDDFT with current approximations is affected by spurious charge-transfer effects when using extended QM regions in GFP. An alternative is to keep a partition of the system in an active site and an external region but to improve upon the MM treatment. Addressing the potential limitations of a MM description in the computation of excitations is currently an active field of investigation,^{58–61} and the exploration of better embedding schemes together with alternative protonations within wild-type GFP will be the subject of future research.

■ ASSOCIATED CONTENT

S Supporting Information. Mulliken charge analysis, TDDFT basis set convergence, and chromophore bond lengths of the B form computed with different QM/MM approaches. PDB files of the B form and the A chain of the A form including residues and waters within 5 Å of the protein. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: c.filippi@utwente.nl, f.buda@chem.leidenuniv.nl, adalgisa.sincropi@unisi.it.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENT

We acknowledge the support from the Stichting Nationale Computerfaciliteiten (NCF-NWO) for the use of the SARA supercomputer facilities. We also acknowledge computational resources provided by the CASPUR and CINECA computer centres. We thank Riccardo Nifosi for useful discussions.

■ REFERENCES

- (1) Day, R. N.; Davidson, M. W. *Chem. Soc. Rev.* **2009**, *38*, 2887–2921.
- (2) Lippincott-Schwartz, J.; Patterson, G. H. *Trends Cell Biol.* **2009**, *19*, 555–565.
- (3) Tsien, R. Y. *Annu. Rev. Biochem.* **1998**, *67*, 509–544.
- (4) Creemers, T. M. H.; Lock, A. J.; Subramaniam, V.; Jovin, T. M.; Völker, S. *Nat. Struct. Biol.* **1999**, *6*, 557.
- (5) Dong, J.; Solntsev, K. M.; Tolbert, L. M. *J. Am. Chem. Soc.* **2006**, *128*, 12038.
- (6) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668–1688.
- (7) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. J. *Comput. Chem.* **2003**, *24*, 1999–2012.
- (8) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.
- (9) CPMD, v3.11.1, C. (revision a11); IBM Corp.: Endicott, NY, 2008; MPI für Festkörperforschung Stuttgart: Stuttgart, Germany, 2001. <http://www.cpmc.org/>.
- (10) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hünenberger, P. H.; Krueger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. *Biomolecular Simulation: The Gromos96 Manual and User Guide*; Hochschulverlag an der ETH Zurich: Zurich, Switzerland, 1996.
- (11) Laio, A.; VandeVondele, J.; Röthlisberger, U. *J. Chem. Phys.* **2002**, *116*, 6941–6947.
- (12) Laio, A.; VandeVondele, J.; Röthlisberger, U. *J. Phys. Chem. B* **2002**, *106*, 7300–7307.
- (13) Troullier, N.; Martins, J. L. *Phys. Rev. B* **1991**, *43*, 1993–2006.
- (14) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865.
- (15) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1997**, *78*, 1396.
- (16) Frisch, M. J.; et al. *Gaussian 09*, Revision A.1; *Gaussian 09*, Revision A.1; Gaussian Inc.: Wallingford, CT, 2009.
- (17) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3096.
- (18) Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
- (19) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
- (20) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51–57.
- (21) Iikura, H.; Tsuneda, T.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 3540–3544.
- (22) Vydrov, O. A.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 234109.
- (23) Aquilante, F.; Vico, L. D.; Ferré, N.; Ghigo, G.; Malmqvist, P.-Å.; Neogrády, P.; Pedersen, T. B.; Pitoňák, M.; Reiher, M.; Roos, B. O.; Serrano-Andrés, L.; Urban, M.; Veryazov, V.; Lindh, R. *Comput. Chem.* **2010**, *31*, 224–247.
- (24) Ghigo, G.; Roos, B. O.; Malmqvist, P.-Å. *Chem. Phys. Lett.* **2004**, *396*, 142–149.
- (25) Forsberg, N.; Malmqvist, P.-Å. *Chem. Phys. Lett.* **1997**, *274*, 196–204.
- (26) Aquilante, F.; Pedersen, T. B.; Lindh, R. *Theor. Chem. Acc.* **2009**, *124*, 1–10.
- (27) Boström, J.; Delcey, M. G.; Aquilante, F.; Serrano-Andrés, L.; Pedersen, T. B.; Lindh, R. *J. Chem. Theory Comput.* **2010**, *6*, 747–754.
- (28) Ponder, J. W.; Richards, F. M. *J. Comput. Chem.* **1987**, *8*, 1016.
- (29) Wang, J.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049.
- (30) CHAMP is a quantum Monte Carlo program package written by C. J. Umrigar, C. Filippi, and collaborators.
- (31) Burkatzki, M.; Filippi, C.; Dolg, M. *J. Chem. Phys.* **2007**, *126*, 234105.
- (32) We add one s and one p diffuse function on the carbon and the nitrogen using exponents from the aug-cc-pVDZ basis set, taken from EMSL Basis Set Library (<http://bse.pnl.gov>).
- (33) Filippi, C.; Umrigar, C. J. *J. Chem. Phys.* **1996**, *105*, 213–226. As the Jastrow correlation factor, we use the exponential of the sum of three fifth-order polynomials of the electron–nuclear (e–n), the electron–electron (e–e), and the pure three-body mixed e–e and e–n distances, respectively. The Jastrow factor is adapted to deal with pseudo-atoms, and the scaling factor κ is set to 0.60 a.u.
- (34) Schmidt, M. W.; Baldrige, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M., Jr. *J. Comput. Chem.* **1993**, *14*, 1347–1363.
- (35) Filippi, C.; Zaccheddu, M.; Buda, F. *J. Chem. Theory Comput.* **2009**, *5*, 2074–2087.
- (36) Casula, M. *Phys. Rev. B* **2006**, *74*, 161102.
- (37) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33–38.

- (38) Yang, F.; Moss, L. G.; Phillips, G. N. *Nat. Biotechnol.* **1996**, *14*, 1246.
- (39) Berman, H.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.; Weissig, H.; Shindyalov, I.; Bourne, P. *Nucleic Acids Res.* **2000**, *28*, 235–242.
- (40) Nifosí, R.; Tozzini, V. *Proteins: Struct., Funct., Genet.* **2003**, *51*, 378–389.
- (41) Elsliger, M.; Wachter, R.; Hanson, G.; Kallio, K.; Remington, S. *Biochemistry* **1999**, *38*, 5296–5301.
- (42) Shinobu, A.; Palm, G. J.; Schierbeek, A. J.; Agmon, N. *J. Am. Chem. Soc.* **2010**, *132*, 11093.
- (43) Zuev, D.; Bravaya, K. B.; Crawford, T. D.; Lindh, R.; Krylov, A. I. *J. Chem. Phys.* **2011**, *134*, 034310.
- (44) Serrano-Andrés, L.; Merchán, M.; Lindh, R. *J. Chem. Phys.* **2005**, *122*, 104107.
- (45) Marques, M. A. L.; López, X.; Varsano, D.; Castro, A.; Rubio, A. *Phys. Rev. Lett.* **2003**, *90*, 257101.
- (46) Sinicropi, A.; Andruniow, T.; Ferré, N.; Basosi, R.; Olivucci, M. *J. Am. Chem. Soc.* **2005**, *127*, 11534–11535.
- (47) Laino, T.; Nifosí, R.; Tozzini, V. *Chem. Phys.* **2004**, *298*, 17.
- (48) The Molcas code has convergence problems with the use of DFT/PBE while no such difficulties are encountered when DFT/BLYP is employed.
- (49) von Lilienfeld, O. A.; Tavernelli, I.; Röhrlisberger, U.; Sebastiani, D. *Phys. Rev. Lett.* **2004**, *93*, 153004.
- (50) von Lilienfeld, O. A.; Tavernelli, I.; Röhrlisberger, U.; Sebastiani, D. *Phys. Rev. B* **2005**, *71*, 195119.
- (51) López, X. Private communication.
- (52) Nielsen, S. B.; Lapierre, A.; Andersen, J. U.; Pedersen, U. V.; Tomita, S.; Andersen, L. H. *Phys. Rev. Lett.* **2001**, *87*, 228102.
- (53) Forbes, M. W.; Jockusch, R. A. *J. Am. Chem. Soc.* **2009**, *131*, 17038–17039.
- (54) Chingin, K.; Balaboin, R. M.; Frankevich, V.; Barylyuk, K.; Nieckarz, R.; Sagulenko, P.; Zenobi, R. *Int. J. Mass Spectrom.* **2011**, *306*, 241–245.
- (55) Epifanovsky, E.; Polyakov, I.; Grigorenko, B.; Nemukhin, A.; Krylov, A. I. *J. Chem. Theory Comput.* **2009**, *5*, 1895–1906.
- (56) Bravaya, K. B.; Khrenova, M. G.; Grigorenko, B. L.; Nemukhin, A. V.; Krylov, A. I. *J. Phys. Chem. B* **2011**, *115*, 8296.
- (57) Hasegawa, J.-Y.; Fujimoto, K.; Swerts, B.; Miyahara, T.; Nakatsuji, H. *J. Comput. Chem.* **2007**, *28*, 2443.
- (58) Wesolowski, T. A. *Phys. Rev. A* **2008**, *77*, 012504.
- (59) Pereira Gomes, A. S.; Jacob, C. R.; Visscher, L. *Phys. Chem. Chem. Phys.* **2008**, *10*, 5353.
- (60) Söderhjelm, P.; Husberg, C.; Strambi, A.; Olivucci, M.; Ryde, U. *J. Chem. Theory Comput.* **2009**, *5*, 649.
- (61) Olsen, J.; Aidas, K.; Kongsted, J. *J. Chem. Theory Comput.* **2010**, *6*, 3721.

Evaluating London Dispersion Interactions in DFT: A Nonlocal Anisotropic Buckingham–Hirshfeld Model

A. Krishtal,^{*,†,‡} D. Geldof,[†] K. Vanommeslaeghe,[§] C. Van Alsenoy,[†] and P. Geerlings^{||}

[†]Department of Chemistry, University of Antwerp, Universiteitsplein 1, B2610 Antwerp, Belgium

[‡]Fachbereich Chemie, Technische Universität Kaiserslautern, Erwin Schrödinger Straße, D-67663 Kaiserslautern, Germany

[§]Department of Pharmaceutical Sciences, University of Maryland School of Pharmacy, 20 Penn St., HSF II-629, Baltimore, Maryland 21201, United States

^{||}Algemene Chemie, Vrije Universiteit Brussel, Pleinlaan 2, B-1050, Brussels, Belgium

ABSTRACT: In this work, we present a novel model, referred to as BH-DFT-D, for the evaluation of London dispersion, with the purpose to correct the performance of local DFT exchange–correlation functionals for the description of van der Waals interactions. The new BH-DFT-D model combines the equations originally derived by Buckingham [Buckingham, A. D. *Adv. Chem. Phys.* **1967**, *12*, 107] with the definition of distributed multipole polarizability tensors within the Hirshfeld method [Hirshfeld, F.L. *Theor. Chim. Acta* **1977**, *44*, 129], resulting in nonlocal, fully anisotropic expressions. Since no damping function has been introduced yet into the model, it is suitable in its present form for the evaluation of dispersion interactions in van der Waals dimers with no or negligible overlap. The new method is tested for an extended collection of van der Waals dimers against high-level data, where it is found to reproduce interaction energies at the BH-B3LYP-D/aug-cc-pVTZ level with a mean average error (MAE) of 0.20 kcal/mol. Next, development steps of the model will consist of adding a damping function, analytical gradients, and generalization to a supramolecular system.

1. INTRODUCTION

The problem of description of London dispersion in density functional theory (DFT) using (semi) local exchange–correlation functionals is a well-known problem.^{1,2} Since the first diagnostic in 1994,¹ an intense discussion on the origin of the problem and appropriate solutions has been ongoing in the DFT community. Generally, one assumes that the cause lies in the local character of the widely used correlation functionals, which, in contrast to the correlation contribution in post-Hartree–Fock methods such as Møller–Plesset or coupled cluster, only utilize information on the density of the system at one point and are therefore unsuitable for the description of a nonlocal phenomenon such as dispersion. Attempts to introduce nonlocal correlation to DFT, such as the random phase approximation (RPA)^{3,4} or the nonlocal van der Waals functionals,^{5–8} are being investigated, but unfortunately the improvement comes with a significant increase in the computational cost. Since the relatively low computational cost of DFT is one of the major factors responsible for its status as the most widely used quantum chemical method today, a range of more pragmatic approaches has been developed to correct the performance of DFT for dispersion interactions. Part of these methods rely on reparametrization of existing local correlation functionals,^{9–11} motivated by the fact that dispersion is partially included in many functionals and that a suitable reparametrization will allow one to achieve the aspired results more consistently. The drawback of such an approach is that the strong empirical character decreases the reliability. For instance, the performance of the reparametrized functionals often decreases for properties other than the electronic energy. Other attempts are based on adding a correction term, representing the dispersion energy, to the energy calculated using standard DFT methods. Also in this category, one can find highly empirical but computationally

attractive methods,¹² based on parameters fitted to reproduce high-level results, as well as the methods with deeper theoretical foundation but computationally more expensive, where *ab initio* information of the systems is used to evaluate the dispersion energy, such as the static or frequency dependent polarizabilities^{13–18} or the dipole moment of the exchange–correlation hole (XDM).^{19–23} Another noteworthy approach is the adaptation of the symmetry adapted perturbation theory²⁴ to the framework of DFT, i.e., SAPT(DFT).^{25–28} SAPT(DFT) has a significant computational advantage against the highly scaling SAPT as the contribution of intramonomer correlation, already embedded within the Kohn–Sham orbitals, does not need to be evaluated. Although possible to use for the correction of DFT dispersion energies,²⁹ SAPT(DFT) is mostly meant for an evaluation of the total interaction energy. The explicit expression for the repulsive contribution of electron–exchange to the dispersion energy within SAPT(DFT), though rarely calculated fully due to the computational expense, offers a more theoretically attractive alternative to the empirical damping functions used in other methods. SAPT(DFT) does have the disadvantage of requiring explicit separation of the system in two parts, which makes it impossible for application on intramolecular dispersion interactions, such as those occurring, for example, in biomolecules.

The model described here offers a compromise between computational simplicity and theoretical rigor by combining expressions derived by Buckingham,³⁰ who used static multipole polarizabilities to express the dispersion interaction between two systems, with the nonempirical Hirshfeld partitioning method. The model offers several advantages: First of all, only *ab initio*

Received: October 11, 2011

Published: November 29, 2011

information about the system is used for the evaluation of the dispersion energy, without making any assumptions about the chemical properties of the system. Second of all, by eliminating frequency-dependence and using *static* atomic polarizabilities, we avoid the necessity to evaluate the computationally expensive Casimir-Polder integrals. Finally, the use of the Hirshfeld method allows us to utilize polarizability *tensors*, which allow one to retain the fully anisotropic character of the dispersion energy.

The model is based on previous work by the present authors,^{17,18,23} where a fully anisotropic model was introduced utilizing the Hirshfeld atomic multipole polarizabilities. The effect of anisotropy was isolated by comparing the results with a fully isotropic model and was found to increase the dispersion energies by up to 30%. In this work, the model is extended in two aspects: First of all, we eliminate the pairwise additivity assumption within the dispersion correction, following Nakai and Sato.¹⁴ They found that eliminating the pairwise additivity assumption in their adaptation of Dobson's local-response model⁵ of frequency-dependent polarizabilities in the evaluation of dispersion interaction reduced the error on C_6 coefficients to only 6%.¹⁴ Second, we introduce two additional higher correction terms into the model, which now includes terms up to R_{AB}^{-10} . The resulting equations are suitable for the evaluation of dispersion energy in van der Waals dimers, with no or negligible overlap between the densities of the monomers. The next development stages of the model will involve introducing a damping function for shorter intermolecular distances where the effect of exchange becomes important, derivation of analytical gradients, and a generalization to intramolecular dispersion interactions.

The full derivation of the new equations is described in section 2, followed by the computational details in section 3 and test results in section 4. A summary and concluding remarks are given in section 5.

2. METHOD

Within the framework of second-order intermolecular perturbation theory, the London dispersion interaction between two systems A and B is defined as the off-diagonal part of the second-order perturbation energy³¹

$$E_{\text{disp}}^{AB} = - \sum_{j_A \neq n_A} \sum_{j_B \neq n_B} \frac{|\langle \psi_{n_A}^{(0)} \psi_{n_B}^{(0)} | \hat{V}^{AB} | \psi_{j_A}^{(0)} \psi_{j_B}^{(0)} \rangle|^2}{(W_{j_A}^{(0)} - W_{n_A}^{(0)}) + (W_{j_B}^{(0)} - W_{n_B}^{(0)})} \quad (1)$$

where $\psi^{(0)}$ and $W^{(0)}$ are the eigenfunctions and eigenvalues of the unperturbed systems, n the ground states, j the excited states, and \hat{V}^{AB} is the electrostatic Coulombic interaction operator of the two systems:

$$\hat{V}^{AB} = \iint \frac{\hat{\rho}_A(\mathbf{r}) \hat{\rho}_B(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \mathrm{d}\mathbf{r} \mathrm{d}\mathbf{r}' \quad (2)$$

In eq 2, $\hat{\rho}(\mathbf{r})$ is the charge density operator, defined as

$$\hat{\rho}(\mathbf{r}) = - \sum_{i=1}^{n_e} \delta(\mathbf{r} - \mathbf{r}_i) + \sum_{j=1}^N Z_j \delta(\mathbf{r} - \mathbf{R}_j) \quad (3)$$

where the first summation is over the n_e electrons of the system and the second summation is over the N nuclei with charges Z . The total charge density is then given by the expectation value of the charge

density operator:

$$\rho(\mathbf{r}) = \langle \psi_n^{(0)} | \hat{\rho}(\mathbf{r}) | \psi_n^{(0)} \rangle \quad (4)$$

This definition of London dispersion assumes a sufficiently large intermolecular distance, such that any electron exchange between the two systems can be neglected.

In order to avoid the pairwise additivity assumption present in our previous model^{17,18,23} as well as in the majority of the DFT-D methods,^{12,15,19–22} the multiatomic character of the systems can be introduced in the beginning of the derivation.³² We will do so by considering the molecular charge density operator $\hat{\rho}_A(\mathbf{r})$ at each point of the space \mathbf{r} as a sum of atomic charge density operators:

$$\hat{\rho}_A(\mathbf{r}) = \sum_{a \in A} \hat{\rho}_a(\mathbf{r}) \quad (5)$$

While the division of the nuclear part of the charge density operator between the atoms is trivial, the definition of the electronic part is not unique. However, such a view of molecular density is common within physical-space atom-in-molecule methods that use a fuzzy-atom approach to the definition of atomic density such as the Hirshfeld method³³ or Becke's partitioning method.² Using the atomic density operators, the electrostatic Coulomb interaction operator can be rewritten as a pairwise sum of diatomic operators:

$$\hat{V}^{AB} = \sum_{a \in A} \sum_{b \in B} \hat{V}^{ab} \quad (6)$$

where the diatomic operator is defined as

$$\hat{V}^{ab} = \iint \frac{\hat{\rho}_a(\mathbf{r}) \hat{\rho}_b(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \mathrm{d}\mathbf{r} \mathrm{d}\mathbf{r}' = \int \hat{\rho}_a(\mathbf{r}) \phi_b(\mathbf{r}) \mathrm{d}\mathbf{r} \quad (7)$$

In eq 7, we have defined the atomic potential $\phi_b(\mathbf{r})$ which represents the potential caused at point \mathbf{r} due to the charge distribution of atom b in system B . Following Buckingham's derivation,³⁰ the atomic potential $\phi_b(\mathbf{r})$ can be expanded in a Taylor series around the origin of atom a , \mathbf{O}_a

$$\begin{aligned} \phi_b(\mathbf{r}) &= \phi_b(\mathbf{O}_a) + \sum_i [\nabla_i \phi_b(\mathbf{O}_a)] (\mathbf{r} - \mathbf{O}_a)_i \\ &+ \frac{1}{2} \sum_i \sum_j [\nabla_i \nabla_j \phi_b(\mathbf{O}_a)] (\mathbf{r} - \mathbf{O}_a)_i (\mathbf{r} - \mathbf{O}_a)_j + \dots \end{aligned} \quad (8)$$

where the subscripts i and j represent the vector components in a Cartesian system of axes: x , y , and z . Subsequently, one can expand the factor $(1)/(|\mathbf{O}_a - \mathbf{r}'|)$ in $\phi_b(\mathbf{O}_a)$ around the origin of atom b , \mathbf{O}_b

$$\begin{aligned} \frac{1}{|\mathbf{O}_a - \mathbf{r}'|} &= \frac{1}{|\mathbf{O}_a - \mathbf{O}_b|} + \sum_i \mathbf{T}_{1i}^{ab} (\mathbf{r}' - \mathbf{O}_b)_i \\ &+ \frac{1}{2} \sum_i \sum_j \mathbf{T}_{2ij}^{ab} (\mathbf{r}' - \mathbf{O}_b)_i (\mathbf{r}' - \mathbf{O}_b)_j + \dots \end{aligned} \quad (9)$$

where the elements of the \mathbf{T}_1 and \mathbf{T}_2 tensors represent the first and second derivatives of $R_{ab} = 1/(|\mathbf{O}_a - \mathbf{O}_b|)$ with respect to the components of \mathbf{R}_{ab}

$$\mathbf{T}_{1i}^{ab} = - \frac{\mathbf{R}_{ab,i}}{R_{ab}^3} \quad (10)$$

$$\mathbf{T}_{2ij}^{ab} = \frac{3\mathbf{R}_{ab,i}\mathbf{R}_{ab,j} - \delta_{ij}R_{ab}^2}{R_{ab}^5} \quad (11)$$

The atomic potential $\phi_b(\mathbf{O}_a)$ can thus be expressed in terms of the atomic charge operators $\hat{q}_b = \int \hat{\rho}(\mathbf{r}) \mathrm{d}\mathbf{r}$, atomic dipole moment

operators $\hat{\mu}_i^b = \int (\mathbf{r} - \mathbf{O}_b)_i \hat{\rho}_b(\mathbf{r}) \, d\mathbf{r}$, atomic quadrupole moment operators $\hat{\Theta}_{ij}^b = \int (\mathbf{r} - \mathbf{O}_b)_i (\mathbf{r} - \mathbf{O}_b)_j \hat{\rho}_b(\mathbf{r}) \, d\mathbf{r}$ and so on.

$$\phi_b(\mathbf{O}_a) = \frac{\hat{q}_b}{R_{ab}} + \sum_i \hat{\mu}_i^b T_{1i}^{ab} + \frac{1}{2} \sum_{ij} \hat{\Theta}_{ij}^b T_{2ij}^{ab} + \dots \quad (12)$$

Inserting eqs 8 and 12 into eq 7 and performing the integration over \mathbf{r} finally results in a diatomic interaction operator expressed in terms of atomic charge and multipole moment operators, and the interatomic distance R_{ab}

$$\hat{V}^{ab} = \frac{\hat{q}_a \hat{q}_b}{R_{ab}} + \sum_i T_{1i}^{ab} (\hat{q}_a \hat{\mu}_i^b - \hat{\mu}_i^a \hat{q}_b) + \frac{1}{2} \sum_{ij} T_{2ij}^{ab} (\hat{q}_a \hat{\Theta}_{ij}^b + \hat{\Theta}_{ij}^a \hat{q}_b - \hat{\mu}_i^a \hat{\mu}_j^b) + \dots \quad (13)$$

By inserting the diatomic interaction operator into eq 1, the total dispersion energy can be expressed as a sum of four-centered atomic contributions

$$E_{\text{disp}}^{AB} = \sum_{a, a' \in A} \sum_{b, b' \in B} E_{\text{disp}}^{aa'bb'} \quad (14)$$

where the first nonzero term in the four-centered atomic dispersion energy is

$$\begin{aligned} E_{\text{disp}}^{aa'bb'}(R^{-6}) &= - \sum_{ijkl} T_{2ij}^{ab} T_{2kl}^{a'b'} \sum_{j_A \neq n_A} \sum_{j_B \neq n_B} \\ &\times \frac{\langle \psi_{n_A}^{(0)} \psi_{n_B}^{(0)} | \hat{\mu}_i^a \hat{\mu}_j^b | \psi_{j_A}^{(0)} \psi_{j_B}^{(0)} \rangle \langle \psi_{n_A}^{(0)} \psi_{n_B}^{(0)} | \hat{\mu}_k^{a'} \hat{\mu}_l^{b'} | \psi_{j_A}^{(0)} \psi_{j_B}^{(0)} \rangle}{(W_{j_A}^{(0)} - W_{n_A}^{(0)}) + (W_{j_B}^{(0)} - W_{n_B}^{(0)})} \\ &= - \sum_{ijkl} T_{2ij}^{ab} T_{2kl}^{a'b'} \sum_{j_A \neq n_A} \sum_{j_B \neq n_B} \\ &\times \frac{\langle \psi_{n_A}^{(0)} | \hat{\mu}_i^a | \psi_{j_A}^{(0)} \rangle \langle \psi_{n_B}^{(0)} | \hat{\mu}_j^b | \psi_{j_B}^{(0)} \rangle \langle \psi_{n_A}^{(0)} | \hat{\mu}_k^{a'} | \psi_{j_A}^{(0)} \rangle \langle \psi_{n_B}^{(0)} | \hat{\mu}_l^{b'} | \psi_{j_B}^{(0)} \rangle}{(W_{j_A}^{(0)} - W_{n_A}^{(0)}) + (W_{j_B}^{(0)} - W_{n_B}^{(0)})} \end{aligned} \quad (15)$$

Equation 15 can be connected to the *distributed static atomic polarizabilities* defined, in the framework of perturbation theory, as

$$\alpha_{kl}^{aa'} = 2 \sum_{j_A \neq n_A} \frac{\langle \psi_{n_A}^{(0)} | \hat{\mu}_k^a | \psi_{j_A}^{(0)} \rangle \langle \psi_{j_A}^{(0)} | \hat{\mu}_l^{a'} | \psi_{n_A}^{(0)} \rangle}{W_{j_A}^{(0)} - W_{n_A}^{(0)}} \quad (16)$$

This definition of distributed polarizability is connected to the classic sum-over-states expression that makes use of molecular dipole moment operators $\hat{\mu}$:

$$\alpha_{kl} = 2 \sum_{j_A \neq n_A} \frac{\langle \psi_{n_A}^{(0)} | \hat{\mu}_k | \psi_{j_A}^{(0)} \rangle \langle \psi_{j_A}^{(0)} | \hat{\mu}_l | \psi_{n_A}^{(0)} \rangle}{W_{j_A}^{(0)} - W_{n_A}^{(0)}} \quad (17)$$

Equation 17 is equivalent to the coupled perturbed Hartree–Fock expression

$$\alpha_{kl} = - \text{tr}[\mathbf{H}^{(k)} \mathbf{D}^{(l)}] = - \int \mathbf{r}_k \rho^{(l)}(\mathbf{r}) \, d\mathbf{r} \quad (18)$$

where $\mathbf{H}^{(k)}$ is the dipole moment matrix, $\mathbf{D}^{(l)}$ is the perturbed density matrix, and $\rho^{(l)}(\mathbf{r})$ is the perturbed molecular density.

In order to connect eq 15 to eq 16, the sum in the denominator in eq 15 is replaced, as suggested by Buckingham,³⁰ by a product,

by introducing average excitation energies U :

$$\begin{aligned} &\frac{1}{(W_{j_A}^{(0)} - W_{n_A}^{(0)}) + (W_{j_B}^{(0)} - W_{n_B}^{(0)})} \\ &= \frac{U_A U_B (1 + \Delta)}{(U_A + U_B)(W_{j_A}^{(0)} - W_{n_A}^{(0)})(W_{j_B}^{(0)} - W_{n_B}^{(0)})} \end{aligned} \quad (19)$$

and neglecting Δ which is given by

$$\Delta = \frac{[U_A^{-1} + U_B^{-1}] - [(W_{j_A}^{(0)} - W_{n_A}^{(0)}) + (W_{j_B}^{(0)} - W_{n_B}^{(0)})]}{(W_{j_A}^{(0)} - W_{n_A}^{(0)}) + (W_{j_B}^{(0)} - W_{n_B}^{(0)})} \quad (20)$$

As one can see, the error made by the approximation is minimal when the average excitation energies U are similar to the energies W of the lowest excited states.

The elements of the atomic static distributed polarizability tensors $\alpha_{ij}^{aa'}$ can be obtained by means of the iterative Hirshfeld method.^{33,34} In the Hirshfeld method, the atomic region is defined through a weight function constructed from the promolecular atomic densities $\rho_a^{\text{pro}}(\mathbf{r})$

$$w_a(\mathbf{r}) = \frac{\rho_a^{\text{pro}}(\mathbf{r})}{\sum_{a' \in A} \rho_{a'}^{\text{pro}}(\mathbf{r})} \quad (21)$$

The fuzzy atom model of the Hirshfeld method is therefore in perfect agreement with the expression of the molecular charge density operators in terms of atomic charge density operators introduced at the beginning of the derivation in eq 5. In the iterative Hirshfeld method (H–I), the spherically symmetric promolecular atomic densities, which represent the densities of the (fictive) atoms prior to bonding, are optimized self-consistently to satisfy the constraint that the number of electrons in a promolecular atom is identical to the number of electrons in the atom in the molecule, i.e.:

$$N_a^{\text{pro}} = \int \rho_a^{\text{pro}}(\mathbf{r}) \, d\mathbf{r} = \int w_a(\mathbf{r}) \rho(\mathbf{r}) \, d\mathbf{r} = N_a \quad (22)$$

As a result, the H–I weight function represents the best definition of an atom in molecule, as retrieved through information theory.³⁵ In particular, the atomic region within the H–I method is defined without any use of parameters or any prior knowledge of the system and therefore reflects the specific chemical surroundings of the atom.

Previously, static atomic intrinsic polarizabilities have been defined within the Hirshfeld method³⁶ by Kristhal et al. and used in our previous pairwise additive model for the evaluation of dispersion interactions.¹⁷ We extend here the definition further to *distributed* atomic intrinsic polarizabilities by partitioning the dipole response μ_k to the applied dipole field μ_j between two different atoms a and a'

$$\alpha_{ij}^{aa'} = - \int (\mathbf{r} - \mathbf{O}_a)_i w_a(\mathbf{r}) w_{a'}(\mathbf{r}) \rho^{(j)a'}(\mathbf{r}) \, d\mathbf{r} \quad (23)$$

The partitioning is thus realized by introducing two weight functions into the definition of polarizability in eq 18 and centering the quantities on the origins of the atoms. The use of two weight functions for the partitioning of a property between two atoms has been previously introduced in the framework of the Hirshfeld method for the partitioning of overlap populations,³⁷ MP2 correlation energy,³⁸ and the definition of

atomic density matrices.^{39,40} In eq 23, $\rho^{(j)a'}(\mathbf{r})$ stands for the molecular density perturbed by a dipole field centered on the atom a' . For systems with zero net charge, $\rho^{(j)a'}(\mathbf{r}) = \rho^{(j)}(\mathbf{r})$ due to the origin independence of the dipole moment in neutral systems. However, for higher multipole moments, the perturbed density requires explicit translation to the center of the atom.⁴² Note that our definition of distributed atomic polarizability differs from previous definitions utilized in the context of evaluation of dispersion interactions. For instance, Hättig et al.⁴¹ have used Stone's model for the calculation of topologically partitioned polarizabilities in a multicenter multipole expansion,⁴³ obtained from a partitioning of the molecular volume according to Bader's QTAIM method.⁴⁴ Sato and Nakai,¹³ on the other hand, have developed a distributed frequency dependent polarizability model based on Dobson's⁵ local-approximation model.

The elements of the dipole–quadrupole distributed atomic intrinsic polarizability tensor \mathbf{A} can be partitioned in two ways. On the one hand, one can consider the dipole response μ_i to the applied quadrupole field Θ_{jk}

$$\mathbf{A}_{1,ijk}^{aa'} = - \int (\mathbf{r} - \mathbf{O}_a)_i w_a(\mathbf{r}) w_{a'}(\mathbf{r}) \rho^{(jk)a'}(\mathbf{r}) d\mathbf{r} \quad (24)$$

On the other hand, one can consider the quadrupole response Θ_{ij} to the applied dipole field μ_k

$$\mathbf{A}_{2,ijk}^{aa'} = - \int (\mathbf{r} - \mathbf{O}_a)_i (\mathbf{r} - \mathbf{O}_a)_j w_a(\mathbf{r}) w_{a'}(\mathbf{r}) \rho^{(k)a'}(\mathbf{r}) d\mathbf{r} \quad (25)$$

Due to the origin dependence of the quadrupole moment, the quadrupole field perturbed density in eq 24 is centered on atom a' by translating it as follows:⁴²

$$\rho^{(ij)a'}(\mathbf{r}) = \rho^{(ij)}(\mathbf{r}) - a'_i \rho^{(j)}(\mathbf{r}) - a'_j \rho^{(i)}(\mathbf{r}) \quad (26)$$

The results obtained from the two definitions are not equivalent, and both are integrated within the present dispersion model by partitioning systematically the response on atoms a (or b) and the applied field on atoms a' (or b'). Since the summation in eq 14 is for both a and a' (or b and b') over all atoms in system A (or B), both possibilities are included for each combination of atoms a and a' (or b and b'). Note that following the notation in eqs 24 and 25, \mathbf{A}_1 is a 3×9 tensor and \mathbf{A}_2 is a 9×3 tensor. The quadrupole (\mathbf{C}), octupole (\mathbf{R}), dipole–octupole (\mathbf{E}_1 and \mathbf{E}_2), and quadrupole–octupole (\mathbf{H}_1 and \mathbf{H}_2) polarizabilities used in this work are defined in a similar fashion, as given in eqs 27–32.

$$\mathbf{C}_{ij,kl}^{aa'} = - \int (\mathbf{r} - \mathbf{O}_a)_i (\mathbf{r} - \mathbf{O}_a)_j w_a(\mathbf{r}) w_{a'}(\mathbf{r}) \rho^{(kl)a'}(\mathbf{r}) d\mathbf{r} \quad (27)$$

$$\mathbf{R}_{ijk,lmn}^{aa'} = - \int (\mathbf{r} - \mathbf{O}_a)_i (\mathbf{r} - \mathbf{O}_a)_j (\mathbf{r} - \mathbf{O}_a)_k w_a(\mathbf{r}) w_{a'}(\mathbf{r}) \rho^{(lmn)a'}(\mathbf{r}) d\mathbf{r} \quad (28)$$

$$\mathbf{E}_{1,ijkl}^{aa'} = - \int (\mathbf{r} - \mathbf{O}_a)_i w_a(\mathbf{r}) w_{a'}(\mathbf{r}) \rho^{(jkl)a'}(\mathbf{r}) d\mathbf{r} \quad (29)$$

$$\mathbf{E}_{2,ijk,l}^{aa'} = - \int (\mathbf{r} - \mathbf{O}_a)_i (\mathbf{r} - \mathbf{O}_a)_j (\mathbf{r} - \mathbf{O}_a)_k w_a(\mathbf{r}) w_{a'}(\mathbf{r}) \rho^{(l)a'}(\mathbf{r}) d\mathbf{r} \quad (30)$$

$$\mathbf{H}_{1,ijklm}^{aa'} = - \int (\mathbf{r} - \mathbf{O}_a)_i (\mathbf{r} - \mathbf{O}_a)_j w_a(\mathbf{r}) w_{a'}(\mathbf{r}) \rho^{(klm)a'}(\mathbf{r}) d\mathbf{r} \quad (31)$$

$$\mathbf{H}_{2,ijk,lm}^{aa'} = - \int (\mathbf{r} - \mathbf{O}_a)_i (\mathbf{r} - \mathbf{O}_a)_j (\mathbf{r} - \mathbf{O}_a)_k w_a(\mathbf{r}) w_{a'}(\mathbf{r}) \rho^{(lm)a'}(\mathbf{r}) d\mathbf{r} \quad (32)$$

In eqs 28, 29, and 31, $\rho^{(ijk)a'}(\mathbf{r})$ is translated to the center of atom a' by

$$\rho^{(ijk)a'}(\mathbf{r}) = \rho^{(ijk)}(\mathbf{r}) - a'_i \rho^{(jk)}(\mathbf{r}) - a'_j \rho^{(ik)}(\mathbf{r}) - a'_k \rho^{(ij)}(\mathbf{r}) + a'_i a'_j \rho^{(k)}(\mathbf{r}) + a'_i a'_k \rho^{(j)}(\mathbf{r}) + a'_j a'_k \rho^{(i)}(\mathbf{r}) \quad (33)$$

The first term in the four-centered atomic dispersion energy in eq 15 can be expressed in terms of the Hirshfeld distributed atomic polarizabilities

$$\begin{aligned} E_{\text{disp}}^{aa'bb'}(R^{-6}) &= - \frac{U_A U_B}{U_A + U_B} \sum_{ijkl} T_{ij}^{ab} T_{kl}^{a'b'} \alpha_{ik}^{aa'} \alpha_{jl}^{bb'} \\ &= - \frac{U_A U_B}{U_A + U_B} \text{tr}[\mathbf{T}_2^{ab} \boldsymbol{\alpha}^{bb'} \mathbf{T}_2^{a'b'} \boldsymbol{\alpha}^{aa'\tau}] \end{aligned} \quad (34)$$

where \mathbf{T}_2 and $\boldsymbol{\alpha}$ are 3×3 tensors and τ designates a transposed matrix. Note that the standard R^{-6} dependence of the first term only becomes explicit when $a = a'$ and $b = b'$. However, in the long-range distance, where the intermolecular distance R_{AB} is significantly larger than the interatomic distances $R_{aa'}$ and $R_{bb'}$, the correct R^{-6} asymptotic behavior of the dispersion energy is reproduced. The higher terms of the dispersion energy in this model are given by

$$\begin{aligned} E_{\text{disp}}^{aa'bb'}(R^{-7}) &= - \frac{U_A U_B}{U_A + U_B} \left\{ \frac{1}{2} \text{tr}[\mathbf{T}_2^{ab} \mathbf{A}_2^{bb'} \mathbf{T}_3^{a'b'\tau} \boldsymbol{\alpha}^{aa'\tau}] \right. \\ &\quad - \frac{1}{2} \text{tr}[\mathbf{T}_2^{ab} \boldsymbol{\alpha}^{bb'} \mathbf{T}_3^{a'b'} \mathbf{A}_1^{aa'\tau}] + \frac{1}{2} \text{tr}[\mathbf{T}_3^{ab} \mathbf{A}_2^{bb'} \mathbf{T}_2^{a'b'} \boldsymbol{\alpha}^{aa'\tau}] \\ &\quad \left. - \frac{1}{2} \text{tr}[\mathbf{T}_3^{ab} \mathbf{A}_2^{aa'} \mathbf{T}_2^{a'b'} \boldsymbol{\alpha}^{bb'\tau}] \right\} \end{aligned} \quad (35)$$

$$\begin{aligned} E_{\text{disp}}^{aa'bb'}(R^{-8}) &= - \frac{U_A U_B}{U_A + U_B} \left\{ \frac{1}{4} \text{tr}[\mathbf{T}_3^{ab} \mathbf{C}^{bb'} \mathbf{T}_3^{a'b'\tau} \boldsymbol{\alpha}^{aa'\tau}] \right. \\ &\quad + \frac{1}{4} \text{tr}[\mathbf{T}_3^{ab} \mathbf{C}^{aa'} \mathbf{T}_3^{a'b'\tau} \boldsymbol{\alpha}^{bb'\tau}] - \frac{1}{4} \text{tr}[\mathbf{T}_2^{ab} \mathbf{A}_2^{bb'} \mathbf{T}_{4(2)}^{a'b'} \mathbf{A}_1^{aa'\tau}] \\ &\quad - \frac{1}{4} \text{tr}[\mathbf{T}_3^{ab} \mathbf{A}_1^{bb'} \mathbf{T}_3^{a'b'} \mathbf{A}_1^{aa'\tau}] - \frac{1}{4} \text{tr}[\mathbf{T}_3^{ab} \mathbf{A}_2^{aa'} \mathbf{T}_3^{a'b'} \mathbf{A}_1^{bb'\tau}] \\ &\quad - \frac{1}{4} \text{tr}[\mathbf{T}_{4(2)}^{ab} \mathbf{A}_2^{bb'} \mathbf{T}_2^{a'b'} \mathbf{A}_2^{aa'\tau}] + \frac{1}{6} \text{tr}[\mathbf{T}_2^{ab} \mathbf{E}_1^{bb'} \mathbf{T}_{4(1)}^{a'b'\tau} \boldsymbol{\alpha}^{aa'\tau}] \\ &\quad + \frac{1}{6} \text{tr}[\mathbf{T}_2^{ab} \boldsymbol{\alpha}^{bb'} \mathbf{T}_{4(1)}^{a'b'} \mathbf{E}_1^{aa'\tau}] + \frac{1}{6} \text{tr}[\mathbf{T}_{4(1)}^{ab} \mathbf{E}_2^{bb'} \mathbf{T}_2^{a'b'} \boldsymbol{\alpha}^{aa'\tau}] \\ &\quad \left. + \frac{1}{6} \text{tr}[\mathbf{T}_{4(1)}^{ab} \mathbf{E}_2^{aa'} \mathbf{T}_2^{a'b'} \boldsymbol{\alpha}^{bb'\tau}] \right\} \end{aligned} \quad (36)$$

$$\begin{aligned} E_{\text{disp}}^{aa'bb'}(R^{-9}) &= - \frac{U_A U_B}{U_A + U_B} \left\{ \frac{1}{12} \text{tr}[\mathbf{T}_2^{ab} \mathbf{A}_1^{bb'} \mathbf{T}_5^{a'b'} \mathbf{E}_1^{aa'\tau}] \right. \\ &\quad \left. - \frac{1}{12} \text{tr}[\mathbf{T}_2^{ab} \mathbf{E}_1^{bb'} \mathbf{T}_5^{a'b'} \mathbf{A}_1^{aa'\tau}] + \frac{1}{12} \text{tr}[\mathbf{T}_3^{ab} \mathbf{H}_1^{bb'} \mathbf{T}_{4(1)}^{a'b'\tau} \boldsymbol{\alpha}^{aa'\tau}] \right\} \end{aligned}$$

$$\begin{aligned}
& -\frac{1}{12} \text{tr}[\mathbf{T}_3^{ab} \mathbf{H}_1^{aa'} \mathbf{T}_{4(1)}^{a'b'} \boldsymbol{\alpha}^{bb' r}] + \frac{1}{8} \text{tr}[\mathbf{T}_3^{ab} \mathbf{C}^{aa'} \mathbf{T}_{4(2)}^{a'b'} \mathbf{A}_1^{bb' r}] \\
& -\frac{1}{8} \text{tr}[\mathbf{T}_3^{ab} \mathbf{C}^{bb'} \mathbf{T}_{4(2)}^{a'b'} \mathbf{A}_1^{aa' r}] + \frac{1}{12} \text{tr}[\mathbf{T}_{4(1)}^{ab} \mathbf{H}_2^{bb'} \mathbf{T}_3^{a'b'} \boldsymbol{\alpha}^{aa' r}] \\
& -\frac{1}{12} \text{tr}[\mathbf{T}_{4(1)}^{ab} \mathbf{H}_2^{aa'} \mathbf{T}_3^{a'b'} \boldsymbol{\alpha}^{bb' r}] + \frac{1}{12} \text{tr}[\mathbf{T}_{4(1)}^{ab} \mathbf{E}_2^{aa'} \mathbf{T}_3^{a'b'} \mathbf{A}_1^{bb' r}] \\
& -\frac{1}{12} \text{tr}[\mathbf{T}_{4(1)}^{ab} \mathbf{E}_2^{bb'} \mathbf{T}_3^{a'b'} \mathbf{A}_1^{aa' r}] + \frac{1}{8} \text{tr}[\mathbf{T}_{4(2)}^{ab} \mathbf{A}_2^{bb'} \mathbf{T}_3^{a'b'} \mathbf{C}^{aa' r}] \\
& -\frac{1}{8} \text{tr}[\mathbf{T}_{4(2)}^{ab} \mathbf{C}^{bb'} \mathbf{T}_3^{a'b'} \mathbf{A}_2^{aa' r}] + \frac{1}{12} \text{tr}[\mathbf{T}_5^{ab} \mathbf{E}_2^{aa'} \mathbf{T}_2^{a'b'} \mathbf{A}_2^{bb' r}] \\
& -\frac{1}{12} \text{tr}[\mathbf{T}_5^{ab} \mathbf{E}_2^{bb'} \mathbf{T}_2^{a'b'} \mathbf{A}_2^{aa' r}] \} \quad (37)
\end{aligned}$$

$$\begin{aligned}
E_{\text{disp}}^{aa'bb'}(R^{-10}) = & -\frac{U_A U_B}{U_A + U_B} \left\{ \frac{1}{36} \text{tr}[\mathbf{T}_{4(1)}^{ab} \mathbf{R}^{bb'} \mathbf{T}_{4(1)}^{a'b'} \boldsymbol{\alpha}^{aa' r}] \right. \\
& + \frac{1}{36} \text{tr}[\mathbf{T}_{4(1)}^{ab} \mathbf{R}^{aa'} \mathbf{T}_{4(1)}^{a'b'} \boldsymbol{\alpha}^{bb' r}] + \frac{1}{36} \text{tr}[\mathbf{T}_2^{ab} \mathbf{E}_1^{bb'} \mathbf{T}_6^{a'b'} \mathbf{E}_1^{aa' r}] \\
& + \frac{1}{36} \text{tr}[\mathbf{T}_6^{ab} \mathbf{E}_2^{bb'} \mathbf{T}_2^{a'b'} \mathbf{E}_2^{aa' r}] + \frac{1}{36} \text{tr}[\mathbf{T}_{4(1)}^{ab} \mathbf{E}_2^{bb'} \mathbf{T}_{4(1)}^{a'b'} \mathbf{E}_1^{aa' r}] \\
& + \frac{1}{36} \text{tr}[\mathbf{T}_{4(1)}^{ab} \mathbf{E}_2^{aa'} \mathbf{T}_{4(1)}^{a'b'} \mathbf{E}_1^{bb' r}] + \frac{1}{24} \text{tr}[\mathbf{T}_3^{ab} \mathbf{C}^{bb'} \mathbf{T}_5^{a'b'} \mathbf{E}_1^{aa' r}] \\
& + \frac{1}{24} \text{tr}[\mathbf{T}_3^{ab} \mathbf{C}^{aa'} \mathbf{T}_5^{a'b'} \mathbf{E}_1^{bb' r}] + \frac{1}{24} \text{tr}[\mathbf{T}_5^{ab} \mathbf{E}_2^{bb'} \mathbf{T}_3^{a'b'} \mathbf{C}^{aa' r}] \\
& + \frac{1}{24} \text{tr}[\mathbf{T}_5^{ab} \mathbf{E}_2^{aa'} \mathbf{T}_3^{a'b'} \mathbf{C}^{bb' r}] - \frac{1}{24} \text{tr}[\mathbf{T}_3^{ab} \mathbf{H}_1^{bb'} \mathbf{T}_5^{a'b'} \mathbf{A}_1^{aa' r}] \\
& - \frac{1}{24} \text{tr}[\mathbf{T}_3^{ab} \mathbf{H}_1^{aa'} \mathbf{T}_5^{a'b'} \mathbf{A}_1^{bb' r}] - \frac{1}{24} \text{tr}[\mathbf{T}_5^{ab} \mathbf{H}_2^{bb'} \mathbf{T}_3^{a'b'} \mathbf{A}_2^{aa' r}] \\
& - \frac{1}{24} \text{tr}[\mathbf{T}_5^{ab} \mathbf{H}_2^{aa'} \mathbf{T}_3^{a'b'} \mathbf{A}_2^{bb' r}] - \frac{1}{24} \text{tr}[\mathbf{T}_{4(1)}^{ab} \mathbf{H}_2^{bb'} \mathbf{T}_{4(2)}^{a'b'} \mathbf{A}_1^{aa' r}] \\
& - \frac{1}{24} \text{tr}[\mathbf{T}_{4(1)}^{ab} \mathbf{H}_2^{aa'} \mathbf{T}_{4(2)}^{a'b'} \mathbf{A}_1^{bb' r}] - \frac{1}{24} \text{tr}[\mathbf{T}_{4(2)}^{ab} \mathbf{H}_1^{bb'} \mathbf{T}_{4(1)}^{a'b'} \mathbf{A}_2^{aa' r}] \\
& \left. - \frac{1}{24} \text{tr}[\mathbf{T}_{4(2)}^{ab} \mathbf{A}_2^{bb'} \mathbf{T}_{4(1)}^{a'b'} \mathbf{H}_1^{aa' r}] + \frac{1}{16} \text{tr}[\mathbf{T}_{4(2)}^{ab} \mathbf{C}^{bb'} \mathbf{T}_{4(2)}^{a'b'} \mathbf{C}^{aa' r}] \right\} \quad (38)
\end{aligned}$$

In eqs 35–38, \mathbf{T}_3 is a 3×9 tensor, $\mathbf{T}_{4(1)}$ is a 3×27 tensor, $\mathbf{T}_{4(2)}$ is a 9×9 tensor, \mathbf{T}_5 is a 9×27 tensor, and \mathbf{T}_6 is a 27×27 tensor. The use of the tensors, as opposed to isotropic values of the polarizabilities, allows one to retain the full anisotropic character of the equations. The effect of anisotropy was previously shown by us to contribute as much as 30% to the dispersion energy;^{17,18} neglecting of the anisotropy results in a loss of all of the terms involving \mathbf{A} , \mathbf{E} , and \mathbf{H} and also results in lower values of terms involving $\boldsymbol{\alpha}$, \mathbf{C} , and \mathbf{R} .

The average excitation energies U_A and U_B are approximated by

$$U_A = \frac{2}{3} \frac{\sum_{a \in A} \langle \mu_a^2 \rangle_{\text{XDM}}}{\sum_{a \in A} \alpha_a^{\text{iso}}} \quad (39)$$

where $\langle \mu_a^2 \rangle_{\text{XDM}}$ is the expectation value of the square of the atomic exchange-hole dipole moment¹⁹ and α_a^{iso} is the isotropic value of the (local) atomic polarizability,³⁶ both obtained using the iterative Hirshfeld method.³⁴ Note that in our previous pairwise additive model^{17,18,23} the excitation energies were atomic properties, located inside the summation over the atoms. This difference prevents us from making a direct comparison with previously obtained results in order to evaluate the effect of the elimination of the pairwise additivity. The downside of the elimination of the pairwise additivity is the formal $\mathcal{O}(N^4)$ scaling as opposed to the $\mathcal{O}(N^2)$ scaling in the pairwise additive model.

However, the steep R^{-n} dependence of the dispersion energy expression can be exploited in order to reduce the scaling behavior to $\mathcal{O}(N^2)$ in larger systems.

Finally, it should be noted that while in the derivation presented here we have used the primitive form of the quadrupole and octupole moments, an analogue derivation can be made using the traceless definition of these properties, leading to expressions similar to eqs 34–38 and identical numerical results for the dispersion energy.

3. COMPUTATIONAL DETAILS

The geometries of the dimers considered in this work were taken from the S22⁴⁵ and SCAI⁴⁶ data sets as well as from ref 18, where all high level interaction energies were obtained at the CCSD(T)/CBS level. The evaluation of dispersion energies is done for the B3LYP functional using the aug-cc-pVTZ and 6-311++G(2df,p) basis sets. This is achieved in three steps: In the first step, the molecular multipole polarizabilities of the monomers are calculated by performing single point calculations with the multipole fields applied in the positive and negative directions with the strength of 0.0001 au using the B3LYP/aug-cc-pVTZ and B3LYP/6-311++G(2df,p) functional and basis set combinations in the Gaussian 09 program.⁴⁷ Consequently, the first order perturbed density matrices are obtained using the finite field method in the BRABO program.⁴⁹ In the second step, the multipole polarizabilities are partitioned into atomic contributions using the STOCK program.⁴⁸ In the last step, the dispersion energies are calculated using the ATDISP program.²³ The interaction energies at the DFT level are obtained using Gaussian 09 with the B3LYP/aug-cc-pVTZ and B3LYP/6-311++G(2df,p) functional and basis set combinations, utilizing the counterpoise method⁵⁰ for correction of the basis set superposition error.

4. RESULTS AND DISCUSSION

As an initial test of the performance of the proposed model, the interaction energies of 34 van der Waals dimers are compared with data obtained at the CCSD(T)/CBS level of theory. The dimers in the test set are gathered from three different databases, namely the S22 database for benchmark on noncovalent complexes,⁴⁵ the SCAI data set containing amino acid side chain interactions,⁴⁶ and several complexes optimized by the present authors in previous work.¹⁸ Since the present model does not take into account the repulsive effect of exchange at shorter distances and no damping function has been implemented at this point, only dimers with sufficiently large interatomic distances are considered, i.e., only dispersion and/or induction bonded dimers. The evaluation of the performance of the model for systems stabilized by shorter-ranged interactions such as hydrogen bonds is postponed to a later stage.

The total interaction energy is obtained by supplementing the BSSE-corrected DFT interaction energy by the dispersion energy obtained using the BH-DFT-D model. Whereas the electrostatic and induction interactions between the monomers are generally assumed to be modeled correctly by (semi)-local exchange functionals, the performance of the functionals for the description of dispersion interactions varies substantially. In order to prevent double-counting, one wishes to utilize a functional which is quasi “dispersion free”. In our previous work,¹⁸ we have observed that the B3LYP functional consequently predicts repulsive interaction energies for pure dispersion-bonded complexes, while, for example the PBE functional

Table 1. The Interaction Energies of a Set of van der Waals Dimers^a

dimer	CCSD(T)/CBS	aug-cc-pVTZ		6-311++G(2df,p)	
		B3LYP	BH-B3LYP-D	B3LYP	BH-B3LYP-D
He ₂ ^(a)	-0.02	0.04	-0.00	0.05	0.03
He-Ne ^(a)	-0.04	0.04	-0.02	0.04	0.02
He-Ar ^(a)	-0.06	0.07	-0.02	0.08	0.03
Ne ₂ ^(a)	-0.07	0.05	-0.05	0.03	-0.03
Ne-Ar ^(a)	-0.12	0.02	-0.14	0.07	-0.03
Ar ₂ ^(a)	-0.27	0.17	-0.19	0.17	-0.08
He-N ₂ L-shaped ^(a)	-0.04	0.09	-0.01	0.10	0.05
He-N ₂ T-shaped ^(a)	-0.06	0.11	-0.03	0.12	0.06
He-FCl ^(a)	-0.10	0.07	-0.07	0.08	0.01
FCl-He ^(a)	-0.13	0.05	-0.34	0.06	-0.16
Ne-CH ₄ ^(a)	-0.18	0.16	-0.20	0.15	-0.07
CH ₄ -C ₂ H ₄ ^(a)	-0.45	0.40	-0.56	0.39	-0.33
SiH ₄ -CH ₄ ^(a)	-0.82	0.55	-1.03	0.59	-0.47
(OCS) ₂ ^(a)	-1.76	0.76	-1.71	0.74	-1.05
C ₆ H ₆ -CH ₄ ^(b)	-1.50	0.77	-1.72	0.76	-1.12
CH ₄ -CH ₄ ^(b)	-0.53	0.38	-0.49	0.40	-0.19
(C ₂ H ₄) ₂ ^(b)	-1.51	0.49	-1.68	0.51	-0.87
(C ₄ H ₄ N ₂) ₂ ^(b)	-4.42	2.44	-4.52	2.49	-3.83
C ₂ H ₄ -C ₂ H ₂ ^(b)	-1.53	-0.66	-1.94	-0.64	-1.70
C ₆ H ₆ -H ₂ O ^(b)	-3.28	-1.20	-4.03	-1.36	-3.49
C ₆ H ₆ -NH ₃ ^(b)	-2.35	-0.11	-2.67	-0.18	-2.16
C ₆ H ₆ -HCN ^(b)	-4.46	-1.98	-4.86	-2.06	-4.23
(C ₆ H ₆) ₂ parallel-displaced ^(b)	-2.73	3.70	-3.59	3.75	-3.07
(C ₆ H ₆) ₂ T-shaped ^(b)	-2.74	0.98	-2.64	0.96	-1.89
C ₆ H ₆ -C ₈ H ₇ N T-shaped ^(b)	-5.73	-0.55	-5.55	-0.54	-4.61
Ala-Leu ^(c)	-1.07	0.84	-1.21	0.88	-0.66
Ile-Ile ^(c)	-1.24	0.72	-1.21	0.75	-0.75
Ile-Leu ^(c)	-1.39	0.41	-0.90	0.43	-0.65
Leu-Gly ^(c)	-0.77	0.26	-0.48	0.27	-0.27
Leu-Leu ^(c)	-1.62	0.40	-0.93	0.43	-0.72
Leu-Thr ^(c)	-1.09	0.37	-0.85	0.39	-0.60
Met-Met ^(c)	-2.03	1.69	-2.36	1.74	-1.47
Val-Leu ^(c)	-1.08	0.40	-0.84	0.42	-0.57
Val-Val ^(c)	-1.39	0.75	-1.41	0.80	-1.03
MAE			0.20		0.36
MAX			0.86		-1.12
R			0.98		0.97

^aThe side chains of the amino acids are noted by the standard three letter codes. The geometries and reference data are obtained from refs 18^(a), 45^(b), and 46^(c). The dispersion corrected values are calculated using eqs 34–38. All interaction energy values are given in kcal/mol. MAE is the unsigned mean absolute error. MAX is the signed maximal absolute error, and R is the correlation coefficient.

already recovers a part of the dispersion energy. For this reason, all “pure” DFT interaction energies in this work are obtained using the B3LYP functional. In ref 17, we have also shown that the dispersion energy is not sensitive to the nature of the functional: very similar dispersion energies were obtained using B3LYP, PBE, and TPSS functionals. Although the precise equations of the model are different in the current work due to, among other things, elimination of the pairwise additivity, we expect the robustness of the model with respect to the choice of the exchange-correlation functional to remain since it mainly depends on the reproduction of static multipole polarizability values. On the other hand, the effect of the size of the basis set utilized for the

calculation of the pure DFT interaction energies and the multipole polarizabilities is investigated by examining a large (aug-cc-pVTZ) and a medium-sized (6-311++G(2df,p)) basis set.

The interaction energies of the selected dimers are listed in Table 1. The difference between the pure B3LYP interaction energies calculated using the aug-cc-pVTZ and the 6-311++G(2df,p) basis sets is mostly small, with a mean absolute error of 0.03 kcal/mol. The interaction energies obtained with the smaller basis set are mostly slightly higher. The largest deviations are observed for the three induction-bonded dimers C₆H₆-H₂O, C₆H₆-NH₃, and C₆H₆-HCN where the B3LYP interaction energy is already attractive: the values obtained using the 6-311+

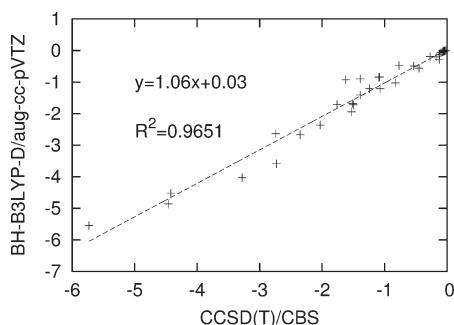


Figure 1. The linear regression parameters for the BH-B3LYP-D/aug-cc-pVTZ model. All values are in kcal/mol.

+G(2df,p) basis set are lower by -0.17 kcal/mol for the $C_6H_6-H_2O$ dimer and -0.08 kcal/mol for the $C_6H_6-NH_3$ and C_6H_6-HCN dimers.

In contrast to the B3LYP interaction energies, the dispersion energies obtained using the two basis sets vary significantly and, as a result, also the total BH-B3LYP-D interaction energy values. The dispersion energies obtained using the smaller basis set are considerably smaller, underestimated by an average 30.2%. This is reflected in the (unsigned) mean absolute error (MAE) and (signed) maximum absolute error (MAX) listed in Table 1 and the linear regression coefficients shown in Figure 1 for the aug-cc-pVTZ basis set and in Figure 2 for the 6-311++G(2df,p) basis set. Despite the underestimation of the dispersion energy values, the regression coefficient for the 6-311++G(2df,p) is very high ($R = 0.97$). The MAE and MAX values, on the other hand, are considerably higher for 6-311++G(2df,p) than for the larger basis set. It should be noted that while the largest error for the 6-311++G(2df,p) basis set is for the $C_6H_6-C_8H_7N$ T-shaped dimer, where the BH-B3LYP-D interaction energy is underestimated by -1.12 kcal/mol compared to the CCSD(T)/CBS value, the largest error for BH-D3LYP/aug-cc-pVTZ is an overestimation of the interaction energy of the $(C_6H_6)_2$ parallel-displaced dimer. The origin for this overestimation lies in the significant contribution of the exchange at the van der Waals minimum: this is reflected in the relative values of the dispersion energy series for this system, where $E_{\text{disp}}(R^{-6}) \approx E_{\text{disp}}(R^{-8})$ (cf. Figure 3 below). This indicates that the series crossed the point of convergence and the exchange-uncorrected multipole expansion becomes invalid. A similar situation is encountered for the induction-bonded $C_2H_4-C_2H_2$ dimer. Other overestimation errors are found in induction-bonded dimers $C_6H_6-H_2O$, $C_6H_6-NH_3$, and C_6H_6-HCN , as well as the Met–Met dimer. Note that for the smaller basis set, the lower pure DFT interaction energies mentioned above for the induction bonded dimers, combined with the smaller dispersion energies values, result in an error compensation for the total BH-B3LYP-D/6-311++G(2df,p) interaction energy. The largest underestimation errors for the aug-cc-pVTZ basis set are observed for the amino acid side chain dimers involving the aliphatic amino acid leucine (Leu), with the largest error found for the Leu–Leu dimer. A possible explanation is the underestimation of the multipole polarizabilities of this amino acid, requiring a larger basis set.

The underestimation of the dispersion energy values calculated with the smaller basis set are directly related to the smaller polarizability values, as can be seen from the molecular isotropic dipole, quadrupole, and octupole polarizability values summarized in Table 2. Herein, the isotropic polarizabilities are defined

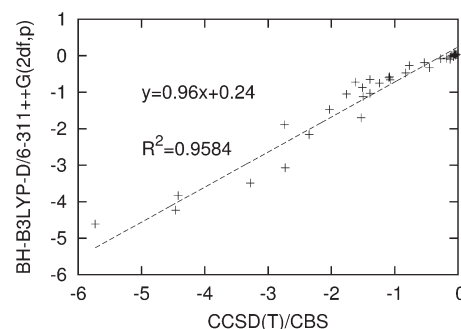


Figure 2. The linear regression parameters for the BH-B3LYP-D/6-311++G(2df,p) model. All values are in kcal/mol.

as $\alpha_{\text{iso}} = 1/3 \sum_i \alpha_{ii}$, $C_{\text{iso}} = 1/9 \sum_{ij} C_{ijij}$, and $R_{\text{iso}} = 1/27 \sum_{ijk} R_{ijkijk}$. One can see that the polarizability values are significantly smaller when calculated using the smaller 6-311++G(2df,p) basis set. The underestimation of values is in particular large for the octupole polarizabilities, where on average only 70% of the value is recuperated using the smaller basis set. It is also noteworthy that the underestimation of the values is larger for the smallest systems, i.e., the rare gas atoms. However, the correlation between the polarizability values obtained using the two basis sets is very high, resulting also in a high correlation between the dispersion energies obtained using the two basis. This indicates that an extrapolation procedure to a complete basis set limit should be possible for the polarizabilities, allowing to obtain more accurate dispersion energies. Alternatively, one can adapt the approach developed by Chong,⁵¹ who constructed specific basis set extensions, with a limited number of additional functions, for the specific goal of obtaining dynamic polarizabilities similar to values obtained at the complete basis set limits. Such procedures are not only desirable for reduction of the computational effort, as calculation of polarizabilities using large basis sets is expensive. They are also desirable in order to avoid numerical stability problems of the finite field procedure used to obtain the perturbed density matrices of the monomers: in larger systems, the convergence of the SCF procedure in the presence of an octupole field can become problematic, as was observed for several dimers from the S22⁴⁵ and SCAI⁴⁶ benchmark sets, excluded from the current study for this reason.

The contributions of the five calculated terms in the dispersion energy series for each examined dimer, calculated at the BH-B3LYP-D/aug-cc-pVTZ level, are summarized in Table 3 and graphically illustrated in Figure 3. In Figure 3, the dimers are ordered as in Table 3 from the lowest to the highest dispersion energies in absolute value and the contribution of each term in the series is displayed in percentage. For the smallest dimers, the $E_{\text{disp}}(R^{-6})$ term accounts for the largest part of the dispersion energy but the contribution of the $E_{\text{disp}}(R^{-8})$ and $E_{\text{disp}}(R^{-10})$ terms increases with the size of the system and the dispersion energy. In the largest systems, the $E_{\text{disp}}(R^{-10})$ contributes as much as 20%, and further study is necessary in order to determine whether this increasing trend will converge. It is possible that a damping function accounting for the effect of exchange on the dispersion energy will reduce the relative contribution of the higher terms. The size of the uneven terms $E_{\text{disp}}(R^{-7})$ and $E_{\text{disp}}(R^{-9})$ fluctuates as it is directly connected to the degree of the anisotropic character in the system: their values are equal to zero for the spherically symmetric rare gas dimers and become increasingly more important in the larger and more anisotropic

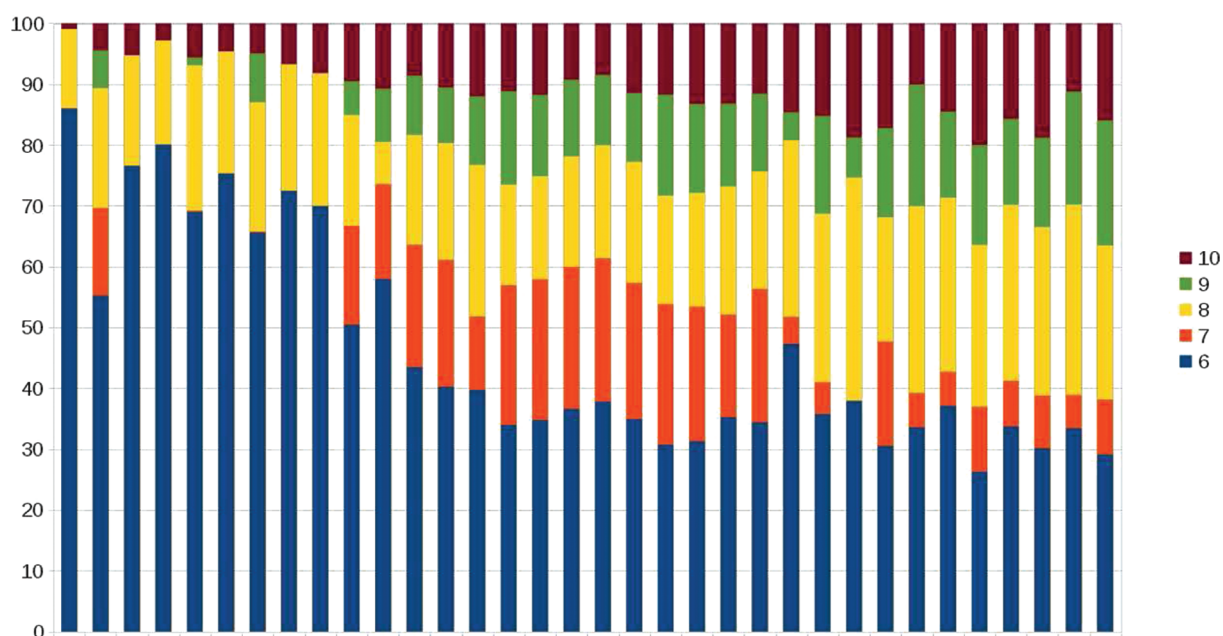


Figure 3. The contribution in percentage of the different terms $E(R^{-n})$, $n = 6-10$, to the total dispersion energy of the examined dimers. The dimers are ordered, as in Table 3, from the smallest to the largest dispersion energy in absolute value.

Table 2. The Isotropic Dipole (α_{iso}), Quadrupole (C_{iso}) and Octupole (R_{iso}) Polarizabilities (in au), Calculated Using the B3LYP Functional and the aug-cc-pVTZ and 6-311++G(2df,p) Basis Sets^a

monomer	α_{iso}		C_{iso}		R_{iso}	
	aug-cc-pVZZ	6-311++G(2df,p)	aug-cc-pVTZ	6-311++G(2df,p)	aug-cc-pVTZ	6-311++G(2df,p)
He	1.47	0.64	0.99	0.43	2.05	0.13
Ne	2.52	1.49	2.91	2.24	6.08	3.04
Ar	11.14	7.89	19.78	15.36	97.27	48.36
N ₂	12.06	11.03	40.91	30.30	145.03	103.21
FCl	18.36	14.73	62.62	45.66	319.62	209.96
CH ₄	17.06	15.33	60.18	43.70	372.62	267.63
C ₂ H ₂	23.87	20.92	96.18	76.10	585.48	392.15
C ₂ H ₄	28.28	26.28	138.99	113.03	1198.09	955.59
C ₆ H ₆	69.53	67.38	658.90	609.72	7801.84	7066.37
C ₄ H ₄ N ₂	59.36	57.72	513.48	472.50	5407.16	4868.19
SiH ₄	32.18	28.65	176.04	130.82	1502.42	1014.32
OCS	34.57	30.83	178.81	142.83	1758.48	1429.94
H ₂ O	9.77	7.60	21.76	14.70	115.41	63.27
NH ₃	14.48	12.01	41.71	29.48	274.68	181.17
C ₈ H ₇ N	103.16	100.64	1392.48	1313.57	29082.92	27025.50
HCN	17.50	15.81	61.75	49.92	434.64	337.03
Ala	29.28	27.45	180.85	152.31	1644.78	1322.04
Leu	66.40	64.62	862.66	807.21	13697.17	12572.92
Ile	65.90	64.13	850.70	796.26	12723.07	11642.16
Gly	17.17	15.32	60.75	44.16	381.30	275.95
Thr	46.64	44.88	439.33	398.42	4932.69	4326.18
Met	76.11	73.39	1046.07	965.80	19445.89	17430.75
Val	53.96	52.34	554.09	510.07	7075.44	6355.31
R	0.9995		0.9997		0.9998	

^aR is the correlation coefficient between the values calculated using the two basis sets.

system where they together contribute up to 40% of the dispersion energy. The relative effect of the anisotropy is expected to remain also

after the inclusion of the damping function. For the most systems, one also observes a converging trend in the dispersion energy series

Table 3. Contribution of the Separate Terms $E_{\text{disp}}(R^{-n})$, $n = 6-10$, to the Dispersion Energy, Calculated Using the BH-B3LYP-D/aug-cc-pVTZ Method (All Values in kcal/mol)

dimer	$E_{\text{disp}}(R^{-6})$	$E_{\text{disp}}(R^{-7})$	$E_{\text{disp}}(R^{-8})$	$E_{\text{disp}}(R^{-9})$	$E_{\text{disp}}(R^{-10})$	E_{disp}
He ₂	-0.03	0.00	0.00	0.00	0.00	-0.04
He-Ne	-0.05	0.00	-0.01	0.00	0.00	-0.06
He-Ar	-0.07	0.00	-0.02	0.00	0.00	-0.09
Ne ₂	-0.07	0.00	-0.02	0.00	0.00	-0.10
He-N ₂ L-shape	-0.06	-0.01	-0.02	-0.01	0.00	-0.10
He-N ₂ T-shaped	-0.09	0.00	-0.03	0.00	-0.01	-0.13
He-FCl	-0.09	0.00	-0.03	-0.01	-0.01	-0.14
Ne-Ar	-0.12	0.00	-0.03	0.00	-0.01	-0.16
FCl-He	-0.19	-0.05	-0.02	-0.03	-0.04	-0.34
Ne-CH ₄	-0.18	-0.06	-0.07	-0.02	-0.03	-0.36
Ar ₂	-0.26	0.00	-0.08	0.00	-0.03	-0.37
Leu-Gly	-0.32	-0.15	-0.13	-0.07	-0.06	-0.74
CH ₄ -CH ₄	-0.35	-0.18	-0.17	-0.08	-0.09	-0.88
CH ₄ -C ₂ H ₄	-0.38	-0.11	-0.24	-0.11	-0.11	-0.95
Leu-Thr	-0.42	-0.28	-0.21	-0.16	-0.14	-1.22
Val-Leu	-0.42	-0.28	-0.21	-0.19	-0.14	-1.24
C ₂ H ₄ -C ₂ H ₂	-0.49	0.01	-0.48	-0.08	-0.24	-1.28
Ile-Leu	-0.48	-0.31	-0.24	-0.16	-0.12	-1.31
Leu-Leu	-0.51	-0.31	-0.25	-0.15	-0.11	-1.34
SiH ₄ -CH ₄	-0.55	-0.35	-0.31	-0.18	-0.18	-1.58
Ile-Ile	-0.59	-0.44	-0.34	-0.32	-0.23	-1.92
Ala-Leu	-0.64	-0.45	-0.38	-0.30	-0.27	-2.05
Val-Val	-0.76	-0.36	-0.46	-0.29	-0.29	-2.16
(C ₂ H ₄) ₂	-0.75	-0.48	-0.42	-0.28	-0.25	-2.18
(OCS) ₂	-1.17	-0.11	-0.72	-0.11	-0.36	-2.47
C ₆ H ₆ -CH ₄	-0.89	-0.13	-0.69	-0.40	-0.38	-2.50
C ₆ H ₆ -NH ₃	-0.95	-0.14	-0.73	-0.36	-0.37	-2.56
C ₆ H ₆ -H ₂ O	-0.96	-0.21	-0.82	-0.40	-0.45	-2.84
C ₆ H ₆ -HCN	-0.96	-0.16	-0.90	-0.53	-0.32	-2.87
(C ₆ H ₆) ₂ T-shaped	-1.21	-0.20	-1.11	-0.72	-0.36	-3.62
Met-Met	-1.24	-0.70	-0.83	-0.59	-0.70	-4.05
C ₆ H ₆ -C ₈ H ₇ N T-shaped	-1.45	-0.45	-1.27	-1.03	-0.80	-5.00
(C ₄ H ₄ N ₂) ₂	-2.10	-0.60	-1.93	-1.02	-1.31	-6.96
(C ₆ H ₆) ₂ parallel-displaced	-1.91	-0.78	-1.94	-1.20	-1.46	-7.29

at the van der Waals minimum, where $E_{\text{disp}}(R^{-6}) > E_{\text{disp}}(R^{-8}) > E_{\text{disp}}(R^{-10})$, indicating that the effect of exchange at these geometries is not dominant, with the exception of the (C₆H₆)₂ parallel-displaced and the C₂H₄-C₂H₂ dimers, as discussed above.

5. SUMMARY AND CONCLUSIONS

In conclusion, we have presented a new combined Buckingham-Hirshfeld model (BH-DFT-D) for the evaluation of dispersion energies at the DFT level. By introducing a pairwise additive Coulomb interaction operator at the beginning of the derivation and defining distributed static multipole polarizability tensors within the framework of the iterative Hirshfeld method,^{33,34,36} a four-centered dispersion energy correction to the interaction energy is obtained. The model constitutes, as such, an improvement on our previous work,^{17,18,23} due to the elimination of the pairwise additivity assumption of the dispersion energy. Furthermore, by making use of atomic polarizability tensors obtained from the *ab initio* molecular polarizabilities of the monomers, the full anisotropic character of the dispersion interaction is preserved. The model in its present form is

suitable for application on dimers with no or only negligible overlap between the densities of the monomers.

The model is tested for a collection of 34 van der Waals dimers at equilibrium geometries by comparing the BH-B3LYP-D interaction energies with high level data obtained at the CCSD-(T)/CBS level. The results obtained using the aug-cc-pVTZ basis set are found to be in good agreement with high level data, with a mean absolute error of 0.20 kcal/mol. The results obtained using the 6-311++G(2df,p) in general underestimate the dispersion energy due to the underestimation of the multipole polarizabilities of the monomers, in particular the octupole polarizability. The results can be further improved by extrapolation of the polarizability values to the complete basis set limit. The effect of the anisotropic terms $E(R^{-7})$ and $E(R^{-9})$ is found to be of increasing importance in the larger systems, contributing as much as 40% to the total dispersion energy value. The necessity of a damping function to compensate for the repulsive contribution of electron exchange to the dispersion energy is observed for several dimers, where the multipole expansion of the Coulomb

operator becomes divergent already at the van der Waals minimum. The next development steps of the model will consist of designing a damping function for shorter distances, derivation, or analytical gradients and generalization of the model to intramolecular dispersion interactions in macromolecules.

AUTHOR INFORMATION

Corresponding Author

*E-mail: alisa.krishtal@ua.ac.be.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

This work was carried out using the Turing HPC infrastructure at the CalcUA core facility of the University of Antwerp, a division of the Flemish Supercomputer Center VSC, funded by the Hercules Foundation, the Flemish Government (department EWI), and the Universiteit Antwerpen. A.K. is grateful to the Research Foundation—Flanders (FWO) for a postdoctoral position and financial support.

REFERENCES

- (1) Kristyán, S.; Pulay, P. *Chem. Phys. Lett.* **1994**, *229*, 175.
- (2) Pérez-Jordá, J. M.; Becke, A. D. *Chem. Phys. Lett.* **1995**, *233*, 134.
- (3) Lotrich, V.; Bartlett, R. J. *J. Chem. Phys.* **2011**, *134*, 184108.
- (4) Eshuis, H.; Furche, F. *J. Phys. Chem. Lett.* **2011**, *2*, 983–989.
- (5) Dobson, J. F.; Dinte, B. P. *Phys. Rev. Lett.* **1996**, *76*, 1780–1783.
- (6) Dion, M.; Rydberg, H.; Schröder, E.; Langreth, D. C.; Lundqvist, B. I. *Phys. Rev. Lett.* **2004**, *92*, 246401.
- (7) Vydrov, O. A.; Voorhis, T. V. *Phys. Rev. Lett.* **2009**, *103*, 063004.
- (8) Lee, K.; Murray, É. D.; Kong, L.; Lundqvist, B. I.; Langreth, D. C. *Phys. Rev. B* **2010**, *82*, 092202.
- (9) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215–241.
- (10) Xu, X.; Goddard, W. A., III. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 2673–2677.
- (11) Zhang, Y.; Vela, A.; Salahub, D. R. *Theor. Chem. Acc.* **2007**, *118*, 693.
- (12) Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1463–1473.
- (13) Sato, T.; Nakai, H. *J. Chem. Phys.* **2009**, *131*, 224104.
- (14) Sato, T.; Nakai, H. *J. Chem. Phys.* **2010**, *133*, 194101.
- (15) Tkatchenko, A.; Scheffler, M. *Phys. Rev. Lett.* **2009**, *102*, 073005.
- (16) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *J. Chem. Phys.* **2010**, *132*, 154104.
- (17) Krishtal, A.; Vannomeslaeghe, K.; Geldof, D.; Van Alsenoy, C.; Geerlings, P. *Phys. Rev. A* **2011**, *83*, 024501.
- (18) Krishtal, A.; Vanommeslaeghe, K.; Olasz, A.; Veszprémi, T.; Van Alsenoy, C.; Geerlings, P. *J. Chem. Phys.* **2009**, *130*, 174101.
- (19) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2007**, *127*, 154108.
- (20) Kannemann, F. O.; Becke, A. D. *J. Chem. Theory Comput.* **2010**, *6*, 1081–1088.
- (21) Steinmann, S. N.; Corminboeuf, C. *J. Chem. Theory Comput.* **2010**, *6*, 1990–2001.
- (22) Steinmann, S. N.; Corminboeuf, C. *J. Chem. Phys.* **2011**, *134*, 044117.
- (23) Olasz, A.; Vanommeslaeghe, K.; Krishtal, A.; Veszprémi, T.; Van Alsenoy, C.; Geerlings, P. *J. Chem. Phys.* **2007**, *127*, 224105.
- (24) Jeziorski, B.; Moszynski, R.; Szalewicz, K. *Chem. Rev.* **1994**, *94*, 1887–1930.
- (25) Heßelmann, A.; Jansen, G. *Chem. Phys. Lett.* **2003**, *367*, 778–784.
- (26) Heßelmann, A.; Jansen, G.; Schtz, M. *J. Chem. Phys.* **2005**, *122*, 014103.
- (27) Misquitta, A. J.; Szalewicz, K. *J. Chem. Phys.* **2005**, *122*, 214109.
- (28) Misquitta, A. J.; Podeszwa, R.; Jeziorski, B.; Szalewicz, K. *J. Chem. Phys.* **2005**, *123*, 214103.
- (29) Rajchel, L.; Żuchowski, P. S.; Szczesniak, M. M.; Chałasinski, G. *Phys. Rev. Lett.* **2010**, *104*, 163001.
- (30) Buckingham, A. D. *Adv. Chem. Phys.* **1967**, *12*, 107.
- (31) London, F. *Trans. Faraday Soc.* **1937**, *33*, 8.
- (32) Stone, A. J.; Tong, C. S. *Chem. Phys.* **1989**, *137*, 121–135.
- (33) Hirshfeld, F. L. *Theor. Chim. Acta* **1977**, *44*, 129–139.
- (34) Bultinck, P.; Van Alsenoy, C.; Ayers, P. W.; Carbo-Dorca, R. *J. Chem. Phys.* **2007**, *126*, 144111.
- (35) Nalewajski, R. F.; Broniatowska, E. *Int. J. Quantum Chem.* **2005**, *101*, 3957.
- (36) Krishtal, A.; Senet, P.; Yang, M.; Van Alsenoy, C. *J. Chem. Phys.* **2006**, *125*, 034312.
- (37) Oláh, J.; Blockhuys, F.; Veszprémi, T.; Van Alsenoy, C. *Eur. J. Inorg. Chem.* **2006**, 69–77.
- (38) Krishtal, A.; Vyboishchikov, S.; Van Alsenoy, C. *J. Chem. Theory Comput.* **2011**, *7*, 2049–2058.
- (39) Vanfleteren, D.; Van Neck, D.; Bultinck, P.; Ayers, P.; Waroquier, M. *J. Chem. Phys.* **2010**, *132*, 164111.
- (40) Vanfleteren, D.; Van Neck, D.; Bultinck, P.; Ayers, P.; Waroquier, M. *J. Chem. Phys.* **2011**, *133*, 231103.
- (41) Hättig, C.; Jansen, G. H.; Hess, B. A.; Ángyán, J. G. *Mol. Phys.* **1997**, *91*, 145–160.
- (42) Geldof, D.; Krishtal, A.; Geerlings, P.; Van Alsenoy, C. *J. Chem. Phys. A* **2011**, *115*, 13096–13103.
- (43) Le Sueur, C. R.; Stone, A. J. *Mol. Phys.* **1993**, *78*, 1267.
- (44) Bader, R. F. W. *Chem. Rev.* **1991**, *91*, 893.
- (45) Jurecka, P.; Sponer, J.; Cerny, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985–1993.
- (46) Berka, K.; Laskowski, R.; Riley, K. E.; Hobza, P.; Vondrasek, J. *J. Chem. Theory Comput.* **2009**, *5*, 982–992.
- (47) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, Revision A.1; Gaussian, Inc.: Wallingford, CT, 2009.
- (48) Rousseau, B.; Peeters, A.; Van Alsenoy, C. *Chem. Phys. Lett.* **2000**, *324*, 189.
- (49) Van Alsenoy, C.; Peeters, A. *THEOCHEM* **1993**, *105*, 19.
- (50) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553.
- (51) Chong, D. P. *Mol. Phys.* **2005**, *103*, 749.

Peculiar Transformations in the $C_xH_xP_{4-x}$ ($x = 0-4$) Series

Alexander S. Ivanov,^{†,‡} Konstantin V. Bozhenko,^{*,†} and Alexander I. Boldyrev^{*,‡}[†]Department of Physical and Colloid Chemistry, Peoples' Friendship University of Russia, 6 Miklukho-Maklaya St., Moscow 117198, Russian Federation[‡]Department of Chemistry and Biochemistry, Utah State University, 0300 Old Main Hill, Logan, Utah 84322, United States Supporting Information

ABSTRACT: In the current work, we performed a systematic study of the $C_xH_xP_{4-x}$ ($x = 0-4$) series using an unbiased CK global minimum and low-lying isomers search for the singlet and triplet $P_4-C_4H_4$ species at the B3LYP/6-31G** level of theory. The selected lowest isomers were recalculated at the CCSD(T)/CBS//B3LYP/6-311++G** level of theory. We found that the transition from a three-dimensional tetrahedron-like structure to a planar structure occurs at $x = 3$, where planar isomers become much more stable than the tetrahedral structures due to significantly stronger π bonds between carbon atoms in addition to increasing strain energy at the carbon atom in the tetrahedral environment.

1. INTRODUCTION

Benzene is by far the most stable isomer for C_6H_6 stoichiometry, with benzvalene and prismane being more than 70 kcal/mol higher in energy.¹ A valence isoelectronic hexaphosphabenzene, on the other hand, is not planar in its most stable benzvalene-like structure. It was recently shown that the transition from the three-dimensional benzvalene-like structure to the planar benzene-like structure in the $C_xH_xP_{6-x}$ ($x = 0-6$) series occurs at $x = 4$.² In our current investigation, we analyze structural transformations in the $C_xH_xP_{4-x}$ ($x = 0-4$) series upon the substitution of a phosphorus atom by the valence isoelectronic C-H group. We demonstrated that P_4 and CHP_3 possess the tetrahedron-like global minimum structures. For the $C_2H_2P_2$ stoichiometry, we found the two most stable structures: derivative of triafulvene and tetrahedron-like, being almost degenerate. For the C_3H_3P and C_4H_4 stoichiometries, we determined that the global minimum structures are vinylacetylene-like. Thus, the 3D-2D transition in the considered series occurs at $x = 3$. We believe that stronger π bonds between carbon atoms as well as increasing strain energy are responsible for this 3D-2D transition.

2. THEORETICAL METHODS AND COMPUTATIONAL DETAILS

A computational search for the global minima structures of P_4 , CHP_3 , $C_2H_2P_2$, C_3H_3P , and C_4H_4 stoichiometries with singlet and triplet electronic states was performed using the Coalescence Kick (CK) program written by Averkiev.³ In the CK method, a random structure is first checked for connectivity: if all atoms in the structure belong to one fragment, then the structure is considered as connected, and the Bery algorithm⁴ for geometry optimization procedure is applied to it. However, in most cases, a randomly generated structure is fragmented; that is, the structure contains several fragments nonbonded with each other including cases with just one atom not being connected. In these cases, the coalescence procedure is applied to the fragmented structure—all of the fragments are pushed to the center of mass simultaneously. The magnitude

of shift should be small enough so that atoms do not approach each other too closely but large enough so that the procedure converges in a reasonable amount of time. In the current version of the CK program, a 0.2 Å shift is used. The obtained structure is checked for connectivity again, and the procedure repeats. When two fragments approach each other close enough, they “coalesce” to form a new fragment, which will be pushed as a whole in the following steps. Obviously, at some point, all fragments are coalesced. This method does not deal with cases when, in a randomly generated structure, two atoms are too close to each other. To avoid this problem, the initial structures are generated in a very large box with all three linear dimensions being 4^* (the sum of atomic covalent radii). Hence, usually an initially generated random structure consists of separated atoms as initial fragments. The current version of the program is designed for the global minimum searches of both single molecules of desired composition and complexes of molecules like solvated anions (e.g., $SO_4^{2-} \cdot 4H_2O$),^{5,6} where the initial geometry of each molecular unit is specified in the input file. In the latter case, the two molecular units of the complex are considered as connected in a fragment if the distances between two of their atoms are less than the sum of the corresponding van der Waals radii.

The CK calculations were performed at the B3LYP level of theory⁷⁻⁹ using the 6-31G** split-valence basis set.¹⁰ Low-lying isomers were reoptimized with followup frequency calculations at the B3LYP level of theory using the 6-311++G** basis set.¹¹⁻¹⁴ Tetrahedron-like and cyclobutadiene-like structures for every stoichiometry were also reoptimized using the CCSD(T) method¹⁵⁻¹⁷ and the 6-311++G** basis set. The final relative energies of the found low-lying isomers were calculated at the CCSD(T)/CBS level by extrapolating CCSD(T)/cc-pvDZ and CCSD(T)/cc-pvTZ¹⁸⁻²² to the infinite basis set using the Truhlar formula.^{23,24} We also calculated relative energies of the two lowest-lying isomers for every stoichiometry at the CCSD(T)/cc-pvQZ level of theory.

Received: October 14, 2011

Published: November 17, 2011

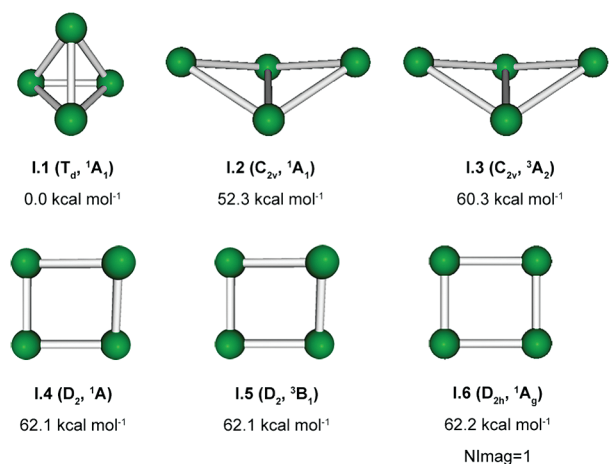


Figure 1. Representative optimized structures of P₄, their point group symmetries, spectroscopic states, and ZPE-corrected (CCSD(T)/6-311+G*) relative energies (CCSD(T)/CBS//CCSD(T)/6-311+G*).

A chemical bonding analysis was performed using Adaptive Natural Density Partitioning (AdNDP)^{25,26} and Natural Bond Orbital (NBO) analysis.^{27,28} The AdNDP approach leads to partitioning of the charge density into elements with the lowest possible number of atomic centers per electron pair: *n*-center–two-electron (nc–2e) bonds, including core electrons, lone pairs, 2c–2e bonds, etc. If some part of the density cannot be localized in this manner, it is represented using completely delocalized objects, similar to canonical MOs, naturally incorporating the idea of the completely delocalized bonding. Thus, AdNDP achieves a seamless description of different types of chemical bonds. The density matrix in the basis of the natural atomic orbitals as well as the transformation between atomic orbital and natural atomic orbital basis sets was generated at the B3LYP/6-31G** level of theory by means of the NBO 3.1 code²⁹ incorporated into Gaussian 09. It is known that the results of the NBO analysis do not generally depend on the quality of the basis set, so the choice of the level of theory for the AdNDP analysis is adequate. All *ab initio* calculations were done using the Gaussian 09 program.³⁰ Molecular structure visualization was performed with the Molden 3.4³¹ and Molekel 5.4.0.8³² programs.

3. RESULTS AND DISCUSSION

P₄ Isomers. The P₄ tetrahedron is known to be a very stable form of phosphorus, and vapor up to 800 °C over phosphorus is completely composed of tetrahedrons.³³ According to our calculations for singlet and triplet states, the singlet P₄ tetrahedral structure is indeed the global minimum structure I.1 (Figure 1) with the butterfly structure I.2 being 52.3 kcal/mol higher (for P₄ structures, relative energies are given at CCSD(T)/CBS//CCSD(T)/6-311+G*). The lowest triplet state isomer I.3 is 60.3 kcal/mol higher than the global minimum. Quasi-planar rectangular tetrphosphacyclobutadiene I.4 was found to be 62.1 kcal/mol higher in energy than the tetrahedral P₄ (Figure 1).

The planar structure I.6 was found to be a minimum at B3LYP/6-311+G* but is a first-order saddle point at CCSD(T)/6-311+G*. Geometry optimization following an imaginary frequency mode leads to the slightly nonplanar structure I.4. Our results are in agreement with calculations reported by Sherer.³⁴

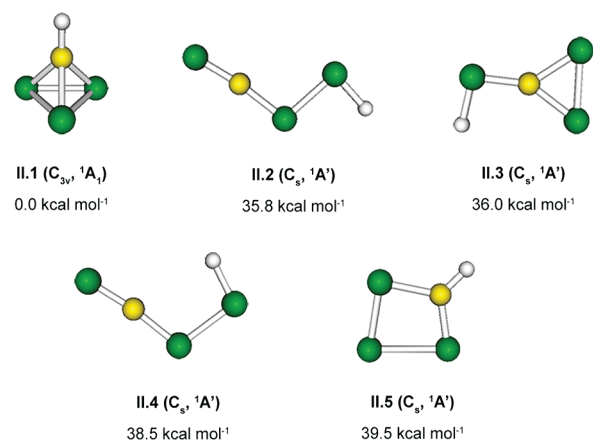


Figure 2. Representative optimized structures of CHP₃, their point group symmetries, spectroscopic states, and ZPE-corrected (B3LYP/6-311++G**) relative energies (CCSD(T)/CBS//B3LYP/6-311++G**).

The only difference is that the planar structure I.6 is not a minimum in our calculations. The distortion of the structure I.6 into the structure I.4 along the a_u imaginary mode occurs due to the pseudo-Jahn–Teller effect (PJT),^{35,36} resulting from vibronic coupling of HOMO–2 (2b_{1u}) and LUMO (1b_{1g}). Indeed, the direct product of their symmetries is the symmetry of the imaginary mode:

$$b_{1u} \otimes b_{1g} = a_u \quad (1)$$

Thus, the symmetry rule³⁷ for the PJT effect is satisfied, as is the second condition:³⁷ the symmetry of the imaginary mode (a_u) of the D_{2h} structure corresponds to the totally symmetric (a) mode in the distorted D₂ isomer. The HOMO–2 and LUMO gap is 8.48 eV (HF/cc-pvTZ//B3LYP/6-311+G*).

CHP₃ Isomers. Our CK global minimum search for the singlet and triplet CHP₃ stoichiometries revealed that the tetrahedral structure II.1 (Figure 2) is the global minimum, and it is significantly more stable than the alternative singlet and triplet structures.

The second lowest isomer, a phosphorus derivative of vinylacetylene (structure II.2), is 35.8 kcal/mol higher in energy than the global minimum structure. The cyclobutadiene-like structure II.5 is 39.5 kcal/mol (here and elsewhere, relative energies of the isomers are given at CCSD(T)/CBS//B3LYP/6-311++G**) higher in energy (Figure 2), but it is now a minimum at both B3LYP/6-311++G** and CCSD(T)/6-311++G**. Apparently, the substitution of one P atom by the C–H group in tetrphosphacyclobutadiene completely quenched the PJT effect. Surprisingly, the UMO–OMO gap in the lowest isomer of CHP₃ that is responsible for out-of-plane distortion is actually slightly smaller (8.34 eV) than the corresponding gap in P₄. Therefore, the simple consideration of PJT may not be always applicable. However, we found in our previous works that the simple PJT consideration worked rather well.^{2,38,39} We were not able to find any theoretical or experimental data regarding CHP₃ isomers in the literature.

C₂H₂P₂ Isomers. The CK search for the global minima of the singlet and triplet C₂H₂P₂ species revealed that the potential energy surface has more low-lying structures than that of P₄ and CHP₃ with the lowest structure III.1 (Figure 3), which can be considered as derivative of triafulvene.

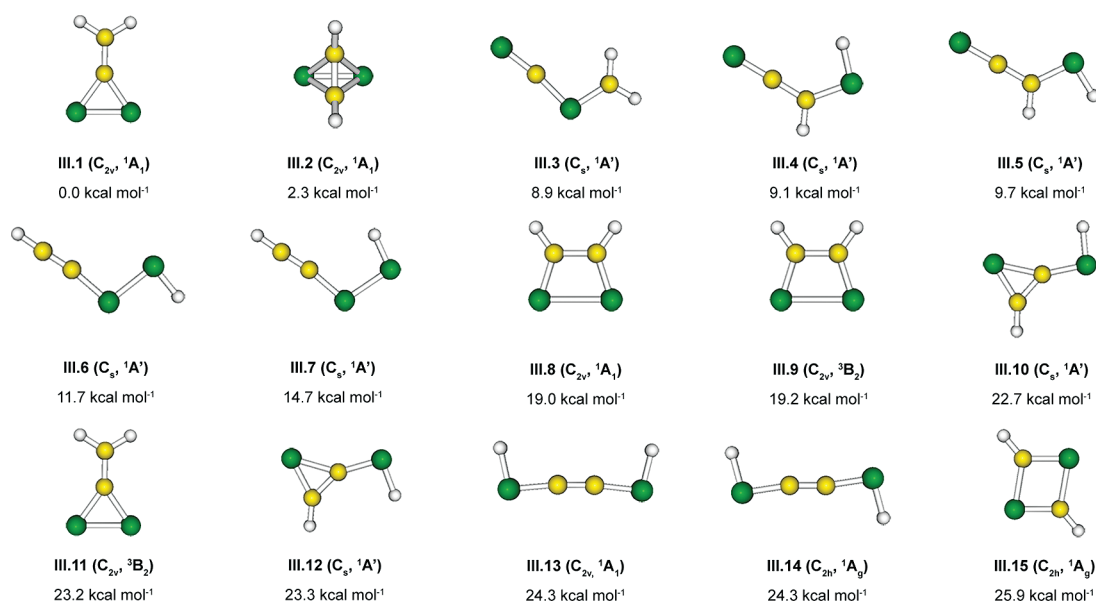


Figure 3. Representative optimized structures of $C_2H_2P_2$, their point group symmetries, spectroscopic states, and ZPE-corrected (B3LYP/6-311++G**) relative energies (CCSD(T)/CBS//B3LYP/6-311++G**).

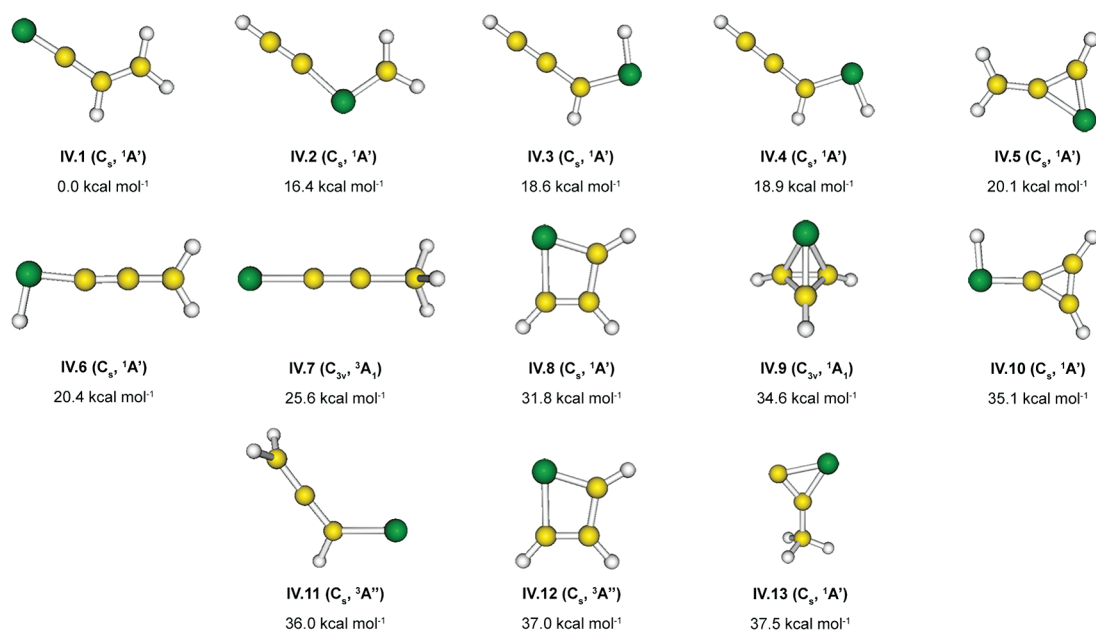


Figure 4. Representative optimized structures of C_3H_3P , their point group symmetries, spectroscopic states, and ZPE-corrected (B3LYP/6-311++G**) relative energies (CCSD(T)/CBS//B3LYP/6-311++G**).

The tetrahedron-like structure III.2 is the second lowest isomer, being just 2.3 kcal/mol higher in energy, and this energy difference is too small to make a definite decision of which of these two structures is the true global minimum. The vinylacetylene-like structures III.3–III.7 are the next set of low-lying isomers, with the relative energies being 8.9–14.7 kcal/mol above the global minimum structure. The *cis*-diphosphacyclobutadiene isomer III.8 is 19.0 kcal/mol higher than isomer III.1 and 16.7 kcal/mol higher in energy than isomer III.2 (Figure 3). The *trans*-diphosphacyclobutadiene isomer III.15 is 25.9 kcal/mol higher than isomer III.1 and 23.6 kcal/mol higher than isomer III.2. The lowest triplet isomer III.9 was found to be 19.2 kcal/mol

higher in energy than the global minimum. Our results are in agreement with previously reported computational data.⁴⁰

C_3H_3P Isomers. From our CK search for the global minimum structure of the singlet and triplet C_3H_3P stoichiometries, we found that the derivative of vinylacetylene (IV.1) is the global minimum (Figure 4), with its other derivatives being the second (IV.2), third (IV.3), and fourth (IV.4) lowest isomers.

The phosphacyclobutadiene isomer IV.8 is more stable than the tetrahedron-like isomer IV.9 by 2.8 kcal/mol (Figure 4). Substitution of three P atoms by three C–H groups in tetraphosphacyclobutadiene switched the relative stabilities of planar and tetrahedron-like isomers. The tetrahedron-like isomer is much

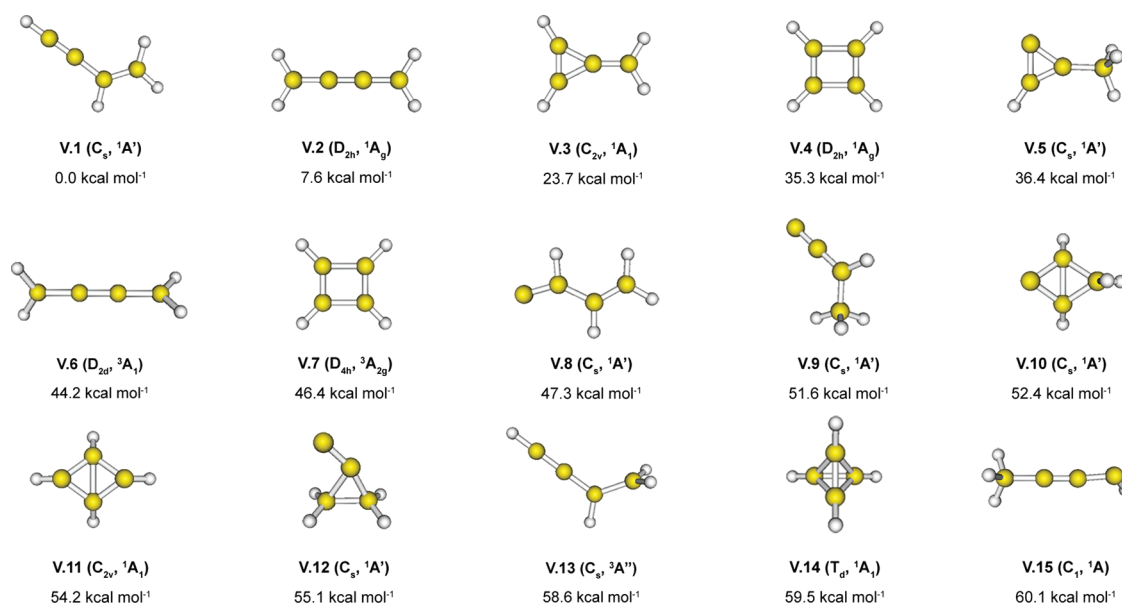


Figure 5. Representative optimized structures of C_4H_4 , their point group symmetries, spectroscopic states, and ZPE-corrected (B3LYP/6-311++G**) relative energies (CCSD(T)/CBS//B3LYP/6-311++G**).

higher (by 34.6 kcal/mol) in energy than the global minimum structure. The lowest triplet isomer IV.7 is 25.6 kcal/mol above the global minimum. Hence, the 3D–2D transition occurs between $C_2H_2P_2$ and C_3H_3P . We are not aware of any experimental or theoretical data for this system.

An opposite 2D–3D transition was observed in the series of mixed boron–aluminum cluster ions, $B_{6-n}Al_n^{2-}$ ($n = 0–6$), and their lithium salts.⁴¹ It was shown that the transition occurs late in the series, at BAl_5^{2-} , and that covalent bonding has an extraordinarily resilient effect that governs the cluster shape more than delocalized bonding does.

C_4H_4 Isomers. The CK search for the global minima revealed that there are a lot of interesting structures on the potential surface of C_4H_4 . The global minimum is well-known vinylacetylene (V.1, Figure 5).

Butatriene and methylenecyclopropene (V.2 and V.3, Figure 5) are 7.6 and 23.7 kcal/mol higher in energy. Our calculations at the CCSD(T)/CBS//B3LYP/6-311++G** level show that cyclobutadiene (V.4) is the fourth lowest isomer and is 24.2 kcal/mol more stable than the tetrahedral structure (V.14). The lowest triplet isomer V.6 was found to be 44.2 kcal/mol higher than vinylacetylene. The triplet aromatic structure V.7 is 11.1 kcal/mol less stable than antiaromatic structure V.4. Our highest-level relative energy values of C_4H_4 isomers are in a good agreement with previous calculations.^{42,43}

Chemical Bonding Pictures Revealed by AdNDP. In order to interpret our results from the chemical bonding point of view, we performed the AdNDP analysis for the global minimum structures. The results of our analysis for these structures are shown in Figure 6.

For the tetrahedral P_4 molecule (I.1), the AdNDP analysis revealed the expected six two-center, two-electron (2c–2e) P–P σ bonds and a lone pair located on each phosphorus atom (not shown in Figure 6), all with occupation numbers (ON) being close to ideal values of 2.00 |e|. The tetrahedron-like structures (II.1 and III.2) have seven 2c–2e (ON = 1.96–2.00 |e|) and eight 2c–2e (ON = 1.97–2.00 |e|)

σ bonds, respectively, as well as a lone pair on each P atom (not shown in Figure 6). For the planar triafulvene-like structure (III.1), we found six 2c–2e (ON = 1.87–2.00 |e|) σ bonds and two 2c–2e (ON = 1.77–2.00 |e|) π bonds. The results of our AdNDP analysis show that the number of π bonds is increasing upon the substitution of P atoms by C–H groups. This has a big influence on the stability of corresponding structures (IV.1, V.1).

According to our systematic computational study, if we consider the relative energies between tetrahedral-like and planar structures along the $C_xH_xP_{4-x}$ ($x = 0–4$) series upon substitution of P atoms by the C–H groups, a transition from the three-dimensional tetrahedron-like structures to the planar structures occurs at $x = 3$. In this case, the increase in relative stability of the planar structure is not related to aromaticity as it was in the $C_xH_xP_{6-x}$ ($x = 0–6$) series, since in our case the planar structures are not aromatic.

There are two main reasons for the switch in the relative stabilities in the considered series along the substitution of phosphorus atoms by the C–H groups. The first one is the strain energy in the tetrahedron structure due to the smaller valence angle at the vertex of the tetrahedron. It is much easier to deform the angle at the phosphorus atom than at the carbon atom. The reported strain energy in tetrahedron C_4H_4 (~119.5 kcal/mol)⁴⁴ is significantly larger than that in tetrahedral P_4 (~14.34 kcal/mol).⁴⁵ Hence, phosphorus structures with acute bonding angles are less strained than carbon analogous structures.

The second factor for the 3D–2D transition is that π -bonding between two carbons is stronger than P–P π -bonding due to a larger overlap of atomic orbitals. The consideration of the $C_xH_xP_{x-4}$ ($x = 0–4$) series proves this statement, because the increase in relative stability of the planar structure is not related to aromaticity as it was in the $C_xH_xP_{6-x}$ ($x = 0–6$) series, since in our case the planar global minimum structures are not aromatic. In the planar vinylacetylene-like molecules, there are two σ bonds less and three π bonds more compared to the tetrahedron-like structures.

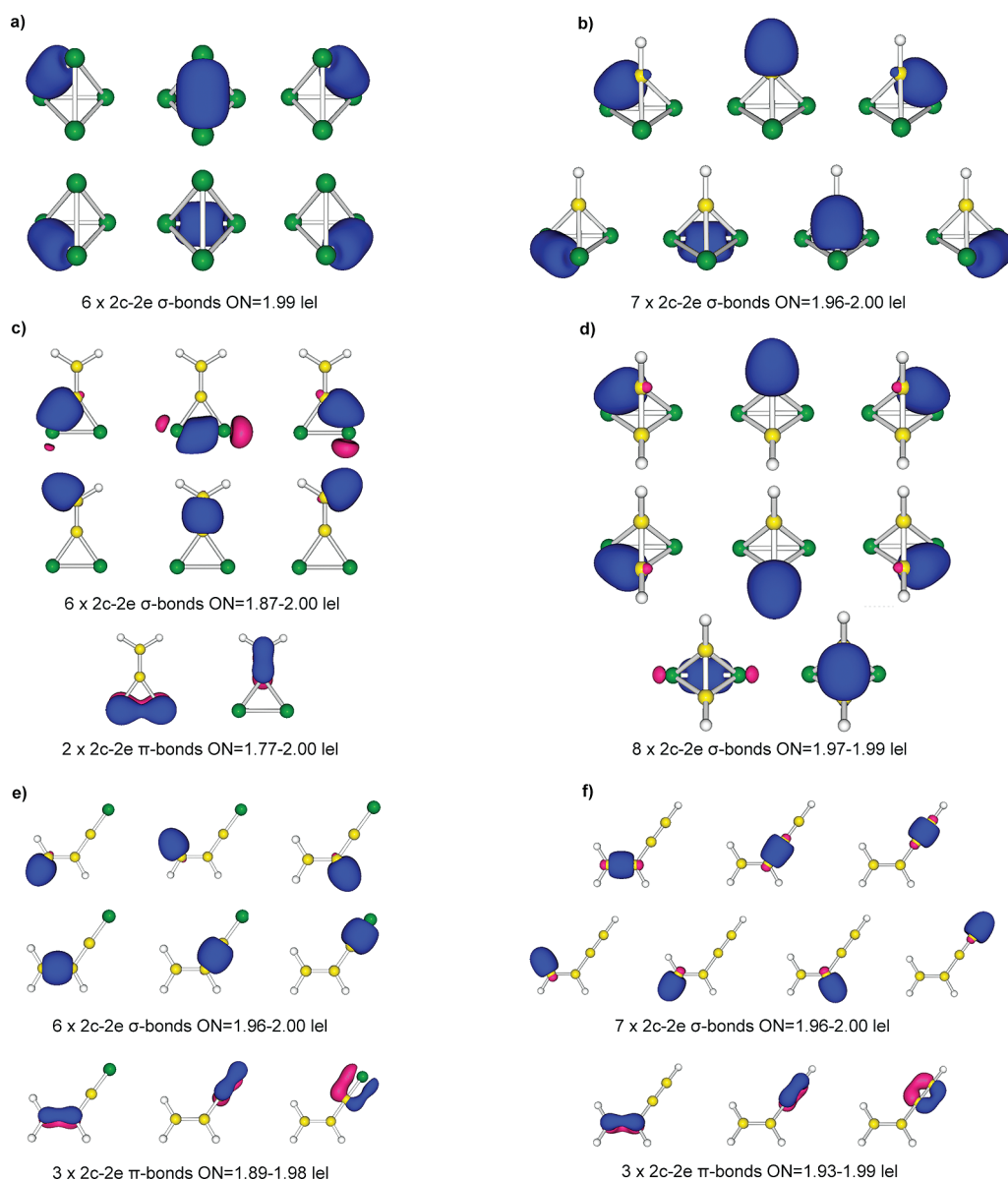


Figure 6. Chemical bonding patterns of the (a) I.1, (b) II.1, (c) III.1, (d) III.2, (e) IV.1, and (f) V.1 global minimum structures revealed by AdNDP.

4. CONCLUSIONS

We presented a systematic study of the $C_xH_xP_{4-x}$ ($x = 0-4$) series. We performed an unbiased CK global minimum and low-lying isomers search for the singlet and triplet $P_4-C_4H_4$ species at the B3LYP/6-31G** level of theory. The selected lowest isomers were recalculated at the CCSD(T)/CBS//B3LYP/6-311++G** level of theory. In addition to that, we calculated relative energies of two isomers at CCSD(T)/cc-pvQZ//B3LYP/6-311++G**. Results at this level of theory are consistent with CCSD(T)/CBS//B3LYP/6-311++G** extrapolations. We found that the global minimum structures and low-lying isomers always have the singlet electronic ground state. The transition from a 3D structure to a 2D structure occurs at $x = 3$ (C_3H_3P), where tetrahedron-like isomers become significantly more unstable than the planar structures.

From the discussion, one can see that six P–P σ bonds in the tetrahedral structure I.1 are much more favorable than four P–P

σ bonds and two P–P π bonds in the quasi-planar structure I.4. Along the series, upon substitution of P atoms by CH groups, the relative stability of tetrahedron-like structures is diminishing, and in C_3H_3P , the planar structure is now more stable than the tetrahedral one, even though the planar molecules are not aromatic. This analysis clearly demonstrates that the 3D–2D transition in the discussed series occurs due to significantly stronger π bonds between carbon atoms in addition to increasing strain energy at the carbon atom in the tetrahedral environment. The importance of π bonding and high strain energy in carbon-rich molecules can also be seen from the global minimum structure IV.1 and structure IV.9, where in the last one the number of σ bonds is increased to three and the number of π bonds is reduced to three. Thus, in evaluating relative stabilities of structures composed of carbon atoms and valence isoelectronic species having phosphorus and silicon, one should keep in mind that carbon-containing molecules would prefer to

maximize the number of π bonds and decrease their bonding strain, whereas in the molecules containing third-row elements, the acute bond angle at the vertex of the tetrahedral structure is the most preferable, and weak π bonds would yield to stronger σ bonds.

■ ASSOCIATED CONTENT

S Supporting Information. Geometry of low-lying structures and their relative energies (CCSD(T)/CBS, CCSD(T)/cc-pvTZ, CCSD(T)/cc-pvDZ and B3LYP/6-311++G**). Relative energies of two low-lying isomers for every stoichiometry at CCSD(T)/cc-pvQZ. This information is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: kbogenko@mail.ru, a.i.boldyrev@usu.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENT

This work was supported by the National Science Foundation (CHE-1057746). The computational resource, the Uinta cluster supercomputer, was provided through the National Science Foundation under Grant CTS-0321170 with matching funds provided by Utah State University.

■ REFERENCES

- (1) Dinadayalane, T. C.; Priyakumar, U. D.; Sastry, G. N. *J. Phys. Chem. A* **2004**, *108*, 11433.
- (2) Galeev, T. R.; Boldyrev, A. I. *Phys. Chem. Chem. Phys.* **2011**, *13*, 20549.
- (3) Sergeeva, A. P.; Averkiev, B. B.; Zhai, H. J.; Boldyrev, A. I.; Wang, L. S. *J. Chem. Phys.* **2011**, *134*, 224304.
- (4) Schlegel, H. B. *J. Comput. Chem.* **1982**, *3*, 214.
- (5) Wang, X. B.; Nicholas, J. B.; Wang, L. S. *J. Chem. Phys.* **2000**, *113*, 10837.
- (6) Wang, X. B.; Sergeeva, A. P.; Yang, J.; Xing, X. P.; Boldyrev, A. I.; Wang, L. S. *J. Phys. Chem. A* **2009**, *113*, 5567.
- (7) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
- (8) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200.
- (9) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
- (10) Binkley, J. S.; Pople, J. A.; Hehre, W. J. *J. Am. Chem. Soc.* **1980**, *102*, 939.
- (11) Rassolov, V. A.; Ratner, M. A.; Pople, J. A.; Redfern, P. C.; Curtiss, L. A. *J. Comput. Chem.* **2001**, *22*, 976.
- (12) Gordon, M. S.; Binkley, J. S.; Pople, J. A.; Pietro, W. J.; Hehre, W. J. *J. Am. Chem. Soc.* **1982**, *104*, 2797.
- (13) Pietro, W. J.; Francl, M. M.; Hehre, W. J.; Defrees, D. J.; Pople, J. A.; Binkley, J. S. *J. Am. Chem. Soc.* **1982**, *104*, 5039.
- (14) Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; Schleyer, P. v. R. *J. Comput. Chem.* **1983**, *4*, 294.
- (15) Cizek, J. *Adv. Chem. Phys.* **1969**, *14*, 35.
- (16) Purvis, G. D.; Bartlett, R. J. *J. Chem. Phys.* **1982**, *76*, 1910.
- (17) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479.
- (18) Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1993**, *98*, 1358.
- (19) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796.
- (20) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007.
- (21) Peterson, K. A.; Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1994**, *100*, 7410.
- (22) Wilson, A.; van Mourik, T.; Dunning, T. H., Jr. *THEOCHEM* **1997**, *388*, 339.
- (23) Truhlar, D. G. *Chem. Phys. Lett.* **1998**, *294*, 45.
- (24) Fast, P. L.; Sanchez, M. L.; Truhlar, D. G. *J. Chem. Phys.* **1999**, *111*, 2921.
- (25) Zubarev, D. Yu.; Boldyrev, A. I. *Phys. Chem. Chem. Phys.* **2008**, *10*, 5207.
- (26) Zubarev, D. Yu.; Robertson, N.; Domin, D.; McClean, J.; Wang, J. H.; Lester, W. A.; Whitesides, R.; You, X. Q.; Frenklach, M. *J. Phys. Chem. C* **2010**, *114*, 5429.
- (27) Foster, J. P.; Weinhold, F. *J. Am. Chem. Soc.* **1980**, *102*, 7211.
- (28) Reed, A. E.; Curtiss, L. A.; Weinhold, F. *Chem. Rev.* **1988**, *88*, 899.
- (29) Glendening, E. D.; Reed, A. E.; Carpenter, J. E.; Weinhold, F. *NBO*, version 3.1; TCI, University of Wisconsin: Madison, WI, 1998.
- (30) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, Revision B.01; Gaussian, Inc.: Wallingford, CT, 2010.
- (31) Schaftenaar, G. *MOLDEN*, version 3.4; CAOS/CAMM Center: The Netherlands, 1998.
- (32) Varetto, U. *Molekel*, version 5.4.0.8; Swiss National Supercomputing Centre: Manno, Switzerland, 2009.
- (33) Cotton, F. A.; Wilkinson, G.; Murillo, C. A.; Bochmann, M. *The Group 15 Elements: P, As, Sb, Bi*. In *Advanced Inorganic Chemistry*, 6th ed.; John Wiley & Sons, Inc.: New York, 1999; p 385.
- (34) Sherer, O. J. *Angew. Chem., Int. Ed.* **2000**, *39*, 1029.
- (35) Bersuker, I. B. *Chem. Rev.* **2001**, *101*, 1067.
- (36) Bersuker, I. B. In *The Jahn-Teller Effect*; Cambridge University Press: Cambridge, U.K., 2006.
- (37) Pearson, R. G. *Proc. Natl. Acad. Sci. U.S.A.* **1975**, *72*, 2104.
- (38) Sergeeva, A. P.; Boldyrev, A. I. *Organometallics* **2010**, *29*, 3951.
- (39) Pokhodnya, K.; Olson, C.; Dai, X.; Schulz, D. L.; Boudjouk, P.; Sergeeva, A. P.; Boldyrev, A. I. *J. Chem. Phys.* **2011**, *134*, 014105.
- (40) Holtz, T. *Phosphaalkynes: from Monomers to Polymers*, Ph. D. thesis, Budapest University of Technology and Economics, Budapest, Hungary, 2010.
- (41) Huynh, M. T.; Alexandrova, A. N. *J. Phys. Chem. Lett.* **2011**, *2*, 2046.
- (42) Cremer, D.; Kraka, E.; Joo, H.; Stearns, J.; Zwier, T. *Phys. Chem. Chem. Phys.* **2006**, *8*, 5304.
- (43) Nemirowski, A.; Reisenauer, H. P.; Schreiner, P. R. *Chem.—Eur. J.* **2006**, *12*, 7411.
- (44) Driess, M.; Noth, H. Bonding in P_4 . In *Molecular Clusters of the Main Group Elements*; WILEY-VCH Verlag GmbH&Co. KGaA: Weinheim, Germany, 2004; p 211.
- (45) Ahlrichs, R.; Brode, S.; Ehrhardt, C. *J. Am. Chem. Soc.* **1985**, *107*, 7260.

Advanced Corrections of Hydrogen Bonding and Dispersion for Semiempirical Quantum Mechanical Methods

Jan Řezáč^{*,†} and Pavel Hobza^{†,‡}

[†]Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic and Center for Biomolecules and Complex Molecular Systems, 166 10 Prague, Czech Republic

[‡]Regional Centre of Advanced Technologies and Materials, Department of Physical Chemistry, Palacky University, 771 46 Olomouc, Czech Republic

S Supporting Information

ABSTRACT: Semiempirical quantum mechanical methods with corrections for noncovalent interactions, namely dispersion and hydrogen bonds, reach an accuracy comparable to much more expensive methods while being applicable to very large systems (up to 10 000 atoms). These corrections have been successfully applied in computer-assisted drug design, where they significantly improve the correlation with the experimental data. Despite these successes, there are still several unresolved issues that limit the applicability of these methods. We introduce a new generation of both hydrogen-bonding and dispersion corrections that address these problems, make the method more robust, and improve its accuracy. The hydrogen-bonding correction has been completely redesigned and for the first time can be used for geometry optimization and molecular-dynamics simulations without any limitations, as it and its derivatives have a smooth potential energy surface. The form of this correction is simpler than its predecessors, while the accuracy has been improved. For the dispersion correction, we adopt the latest developments in DFT-D, using the D3 formalism by Grimme. The new corrections have been parametrized on a large set of benchmark data including nonequilibrium geometries, the S66x8 data set. As a result, the newly developed D3H4 correction can accurately describe a wider range of interactions. We have parametrized this correction for the PM6, RM1, OM3, PM3, AM1, and SCC-DFTB methods.

INTRODUCTION

Until very recently, semiempirical quantum mechanical (SQM) methods were developed mainly to reproduce the thermochemical properties of molecules. Because of the many approximations used and since no greater attention has been paid to them, noncovalent interactions have not been described well by the SQM methods. On the other hand, the SQM methods are the most efficient methods that still use a quantum mechanical description of the system, which provides them with several advantages over fully empirical molecular mechanics (MM). For one, the SQM methods are able to describe quantum effects that are not covered by molecular mechanics; another advantage is that SQM methods can be applied to any molecule without previous parametrization. The SQM methods can be derived from the ab initio Hartree–Fock (HF) method by introducing further approximations, such as reduction of the basis set to the absolute minimum and neglect or simplification of a large part of the integrals. To compensate for these approximations, additional empirical terms are added. The parameters, both in the integrals and in the additional corrections, are either derived directly from the experimental data or optimized to reproduce them. Although not in the mainstream, the SQM methods are still evolving, and the recently published PM6,¹ RM1,² OM-x,³ and other methods have brought substantial improvements over their predecessors. When they are combined with linear scaling algorithms, such as the localized orbital method MOZYME,⁴ it is possible to calculate routinely whole proteins at the SQM level.

Two types of interactions that are difficult to describe using the SQM methods are London dispersion and hydrogen bonds (H-bonds), both of which are common in the studied systems

and crucial for obtaining accurate results. The London dispersion can be described explicitly only by methods that account for electron correlation. The SQM methods can include part of this interaction by other means, through the parameters and core–core potentials, but a major part of the dispersion is still missing. In methods where the dispersion is missing completely, such as in the HF or density functional theory (DFT), it can be easily added as an a posteriori empirical correction (the resulting corrected DFT is referred to as DFT-D). A pairwise potential based on the physically sound c_6/r^6 formula (where c_6 is a coefficient determining the strength of the interaction and r is the interatomic distance) scaled at short distances by a damping function has proven to be a very effective solution. It is widely used in DFT and has also been applied to the semiempirical methods,^{3,5} in some cases in conjunction with a partial reparameterization of the SQM method itself.

However, none of the resulting dispersion-corrected SQM methods had been accurate enough to provide a quantitative description of all of the types of noncovalent interactions until corrections for both dispersion and hydrogen bonding were applied. We were the first to develop a H-bond correction for the SQM method,⁶ namely, PM6. In this approach, the correction energy is a function of the hydrogen-bond distance and angle, and of partial charges of the atoms involved in the H-bond. In combination with a parametrization of a dispersion correction used previously in DFT, the resulting method (PM6-DH)

Received: October 25, 2011

Published: December 22, 2011

achieved an accuracy of less than 1 kcal/mol in small model complexes.

The second generation of the correction⁷ (DH2) was developed to solve the problems of the first version. The dispersion correction was modified in order to avoid double counting of the dispersion energy already described by the underlying method. This was achieved by scaling the whole dispersion term and a specific scaling of the c_6 coefficient for sp^3 -hybridized carbon. In the H-bonding correction, discontinuities of the potential were fixed, and additional geometrical parameters of the hydrogen bond were added to avoid false contributions from atoms not involved in a real H-bond and to improve the geometry of the H-bond. The DH2 correction was parametrized for use with multiple semiempirical methods, namely, PM6,¹ AM1,⁸ OM3,³ and SCC-DFTB;⁹ the best results have been achieved with PM6.

The third generation of the correction¹⁰ (DH+) addresses two problems of the DH2 correction. The first is the use of partial atomic charges from the underlying semiempirical calculation in the H-bond energy correction. The derivative of the charge with respect to the coordinates, which is expensive to calculate, enters the expression for the gradient of the correction. For practical purposes, the derivative of the charges was assumed to be zero, but this approximation cannot be used in some cases, such as in accurate optimizations or in molecular dynamics. In the DH+ correction, the charges are no longer used, and the exact gradient can be obtained easily. The second issue addressed is the fixed definition of the hydrogen donor and acceptor atoms. In the DH2 formalism, a proton transfer along a hydrogen bond exhibits a discontinuous PES. In the DH+ correction, the two potentials for both the reactant and product are switched smoothly. The dispersion correction in the DH+ is identical to the one in DH2.

It is clear from this brief review that many problems have been successfully solved during the development of the corrections for the SQM methods. On the other hand, the complexity of the H-bonding correction has grown substantially. The energy of the DH2 and DH+ corrections depends not only on the atom distances and angle of the H-bond but also on multiple other internal coordinates. Not only does this make the actual calculation complicated, the main problem is in defining these coordinates for different groups involved in the H-bond. In practical implementation, this information on all of the possible H-bonds in the system is stored in memory, making the calculation inefficient for large molecules.

Additionally, two more important points were neglected in the construction of the DH+ correction. First, the earlier versions of the H-bonding correction used the atomic charges and thus naturally described strong hydrogen bonds involving charged groups. In DH+, the same parameters are used for neutral and charged H-bonds, which leads to an underestimation of the interaction in charged systems. Second, the angular terms in both DH2 and DH+ do not have smooth first derivatives, which makes it impossible to optimize the geometry of some systems.

Despite the limitations described above, the PM6-DH2 method has been successfully applied to practical problems.^{11–14} It was used for calculations of protein–ligand interactions in computational drug design, yielding much better correlation with the experimental data than an equivalent protocol based on molecular mechanics. The great potential of the corrected SQM methods in this area and other possible applications has led us to develop the corrections further.

Here, we propose the next generation of the H-bonding and dispersion corrections for semiempirical methods. In the H-bond

Table 1. A Comparison of the Hydrogen-Bonding Correction in the DH2, DH+, and Newly Introduced D3H4 Approaches

	H2	H+	H4
exact gradient	NO	YES	YES
proton transfer	NO	YES	YES
accurate for charged systems	YES	NO	YES
smooth energy derivatives	NO	NO	YES
coordinates per H-bond (torsions)	4 (2)	7 (4)	3 (0)

Table 2. A Comparison of the Dispersion Correction in the DH2, DH+, and Newly Introduced D3H4 Approaches

	D2, D+	D3
parameters for elements	18	94
valence-dependent parameters	NO	YES
parameters	element-wise	pairwise

correction, we have wanted to preserve the improvements brought by the DH+ approach, solve its poor performance in charged systems, and, importantly, simplify the form of the correction. Unlike its predecessors, the new correction has not only a smooth potential energy surface but also its first and second derivatives. Another important feature is that the correction potential is strictly local and does not have to be evaluated for more distant potential H-bonds. This makes the computational expense grow only linearly with the size of the calculated system. Finally, our goal is to improve the accuracy, or at least keep it at the level of the previous, more complex approaches. These developments are summarized in Table 1, which lists the most important features of the correction in the DH2, DH+, and D3H4 versions.

We have also updated the dispersion correction, adopting the latest advances in the DFT-D methods.^{15–17} We have based our dispersion correction on the DFT-D3 method.¹⁷ The improvement of the accuracy is not large, but the new approach has other important benefits (see Table 2). First, it uses a large set of atomic parameters consistently constructed for all of the elements up to plutonium. This makes it a useful complement to the PM6 method, which can treat 70 elements; for many of these, the parameters for the earlier dispersion correction were missing. The DFT-D3 correction uses different parameters for the possible valence states of the atoms and switches between them smoothly.

In this work, we have focused mainly on PM6, which we found to be the most accurate SQM method for the description of biomolecular systems. Additionally, we report the parametrization of the correction for many other semiempirical methods we have used in our previous work, namely, AM1⁸ and OM3³ and the self-consistent charge density-functional tight-binding⁹ (SCC-DFTB) method. In this study, we parametrize the corrections for two more methods, PM3¹⁸ and the more recent RM1.²

The dispersion and hydrogen bonding corrections described here are close to the accuracy limit that could be achieved with *a posteriori* corrected semiempirical methods. The D3H4 approach also solves all of the issues we encountered in the previous generations of the corrections. Therefore, we consider the D3H4 corrections to be a final version that can be recommended for general use.

It should also be noted that empirical corrections are not the only way to achieve a better description of hydrogen bonds in semiempirical methods. There is no fundamental reason that would

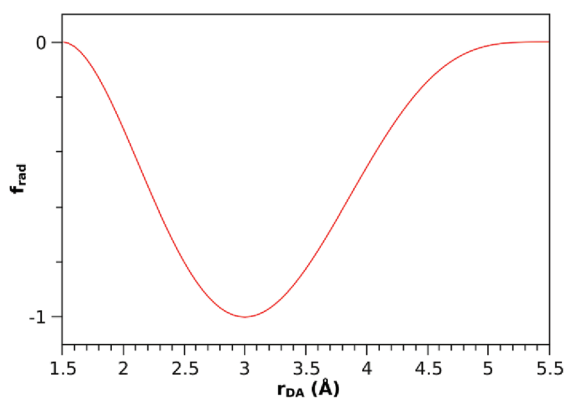


Figure 1. Radial potential of the hydrogen bond correction. This polynomial function is scaled by a coefficient specific for each combination of donor and acceptor elements.

prevent SQM methods from describing hydrogen bonds quantum mechanically. This had been discussed in the literature,^{19,20} and significant improvements have been achieved when polarization functions were added to hydrogen atoms. SINDO1²¹ was the first SQM method using this approach. The recently introduced PMO method²² yields very promising results, but its parametrization is limited to hydrogen and oxygen only.

H-BOND CORRECTION

The correction potential for each potential hydrogen bond is constructed from the radial part f_{rad} , which determines the strength of the correction from the donor–acceptor distance (r_{DA}) scaled by the angular term (f_{ang}), and the proton transfer term (f_{PT}), which depends on the position of the hydrogen between the donor and acceptor. In the case of charged groups, additional scaling by the f_{charge} term is applied to make the correction stronger, and when water acts as a hydrogen donor, further scaling f_{wat} is applied. The complete expression is

$$E_{\text{HB}} = c \times f_{\text{rad}}(r_{\text{DA}}) \times f_{\text{ang}}(\alpha_{\text{DHA}}) \times f_{\text{PT}}(r_{\text{DH}}, r_{\text{AH}}) \times f_{\text{charge}} \times f_{\text{wat}} \quad (1)$$

where c is the parameter determining the strength of the correction, α_{DHA} is the donor–hydrogen–acceptor angle (defined as zero in the linear arrangement), and r_{DH} and r_{AH} are the distances between the hydrogen and the donor and acceptor.

Radial Potential. The shape of this potential mimics the difference between the dissociation curve of the H-bond calculated with the corrected method and a reference. This difference does not vanish even at larger distances (6–10 Å). In the previous versions of the H-bond correction, a $1/r^x$ form damped at short distances was used. The unlimited range of such a correction led to problems in condensed systems with many potential H-bonds in this range. Small contributions that are not a real H-bond added up to an erroneous stabilization. This had been addressed by the addition of further criteria for the identification of true H-bonds based on additional internal coordinates. This approach works well but makes the calculation rather complex. Another solution, developed in this work, is to make the correction potential more short-ranged. We found that H-bonds with a donor–acceptor distance larger than 5.5 Å can be left uncorrected without a loss of accuracy, and we use this cutoff radius in our correction. The correction potential

(Figure 1) is a polynomial determined by the following points: It has a minimum at the average H-bond distance, $r_{\text{DA},0} = 3.0$ Å. At a cutoff radius of 5.5 Å, it smoothly approaches zero. Here, the first and second derivatives are also required to be smooth. The third point defines the curvature of the potential at shorter than equilibrium distances by setting the distance where the correction approaches zero. The distance $r_{\text{DA},\text{min}} = 1.5$ Å was determined by fitting the dissociation curves of the training set. This region of the potential is unlikely to be visited. Therefore, we make sure that the energy surface is smooth, but we do not apply any conditions to the energy derivatives. The eight conditions described here are used to construct a seventh-order polynomial. Outside this interval, the correction function is set to zero (the correction is not calculated in a practical implementation). The obtained coefficients are listed in eq 2 in a rounded form; in a real implementation, it is necessary to use more precise values in order to avoid large errors in the result. The coefficients are provided at high precision in the Supporting Information (Table S4).

$$f_{\text{rad}}(r_{\text{DA}}) = -0.003r_{\text{DA}}^7 + 0.074r_{\text{DA}}^6 - 0.701r_{\text{DA}}^5 + 3.253r_{\text{DA}}^4 - 7.207r_{\text{DA}}^3 + 5.318r_{\text{DA}}^2 + 3.407r_{\text{DA}} - 4.685 \quad (2)$$

The depth of the minimum of this potential is determined by the only free parameter in the correction, coefficient c . The parameters obtained by fitting to reference values (as described below) are tabulated for all of the combinations of donor and acceptor elements.

Angular Term. Goniometric functions, $\cos(\alpha)$ in DH2 and $\cos(\alpha)^2$ in DH+, were used previously, as they seem to be a natural expression of the angular dependence of the potential. However, the derivative of this term has a cusp at $\alpha = 0$; it is in the linear arrangement of the H-bond (Figure 2). This makes it practically impossible to optimize a system with a linear hydrogen bond. For this reason, and in order to gain more control over the shape of the potential, we replaced the cosine with a polynomial constructed to have smooth derivatives. First, we define a polynomial switching function $f_{\text{sw}}(x)$ that smoothly changes from 0 at $x = 0$ to 1 at $x = 1$, having first and second derivatives of zero at the boundaries of this interval.

$$f_{\text{sw}}(x) = -20.0x^7 + 70.0x^6 - 84.0x^5 + 35.0x^4 \quad (3)$$

The angular term then uses this function to construct a potential with the desired properties in the interval of 0 to $\pi/2$:

$$f_{\text{ang}}(\alpha_{\text{DHA}}) = 1 - (f_{\text{sw}}(2\alpha_{\text{DHA}}/\pi))^2 \quad (4)$$

We have tested multiple functions with different shapes, namely, the width of the minimum, and one with a rather wide, flat maximum (eq 4) worked best (in terms of interaction energies in the model complexes).

Proton-Transfer Term. The correction described so far would work on equilibrium geometries but breaks down when the hydrogen is moved along the H-bond to the acceptor. We address this analogously to the H+ correction by scaling the correction using a switching function dependent on the donor–hydrogen distance. When the hydrogen is at a covalent distance, this function should be equal to 1. In contrast to the H+ correction, the coefficient c changes when the donor and acceptor are exchanged during the proton transfer (the donor is defined as the atom closer to the hydrogen). A smooth transition is ensured by the proton transfer switching function, which scales the whole correction to zero when this exchange occurs (at $r_{\text{DH}} = r_{\text{AH}}$). We use the polynomial switching function f_{sw} defined in eq 3. Here, it

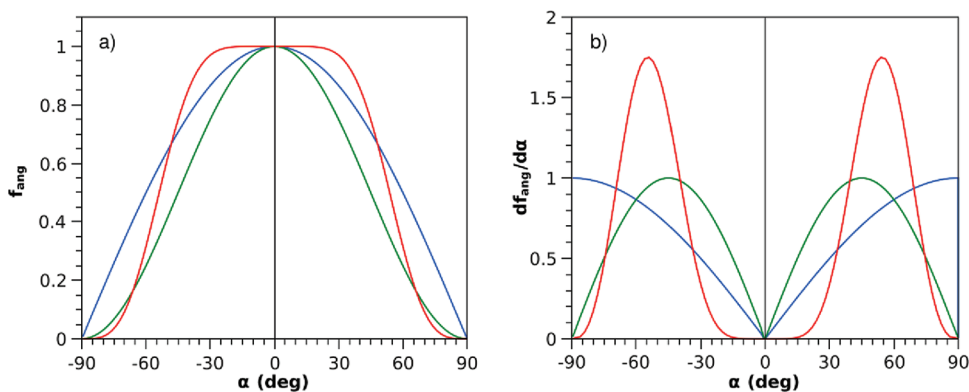


Figure 2. Angular term (a) and its derivative (b) in the DH2 (blue), DH+ (green), and D3H4 (red) hydrogen-bond corrections. The angle of zero degrees corresponds to a linear arrangement of the atoms.

switches from 1 to 0 in the interval from $r_0 = 1.15 \text{ \AA}$ (\sim max. covalent bond length) to $r_1 =$ half of the sum of the donor–hydrogen and acceptor–hydrogen distances.

$$f_{\text{PT}}(r_{\text{DH}}) = \begin{cases} 1 & \text{when } r_{\text{DH}} < r_0 \\ 1 - f_{\text{sw}}((r_{\text{DH}} - r_0)/(r_1 - r_0)) & \text{in } (r_0 \dots r_1) \end{cases} \quad (5)$$

Scaling of Charged H-Bonds. When the donor or/and acceptor are charged, the hydrogen bond becomes stronger than in a neutral system. The first two generations of the correction accounted for this directly, because the atomic charges determined the strength of the correction. In the third generation, these charges were replaced by fixed parameters, which leads to substantial errors for charged H-bonds. Here, we correct this problem by additional scaling of the correction for charged groups. The scaling factors are fitted to the reference data as described below. The identification of the charged groups can be done automatically in simple cases. We implemented it for the COO[−] and NHR3⁺ groups most common in biomolecules. For calculations on the minimum geometry, this scaling can be applied easily on the basis of the atom types determined by a connectivity search. This approach is not valid in reactions, e.g., proton transfer, where atom types defined this way would change abruptly. To keep the potential continuous, we have introduced a fractional measure of atom valence as the description of the similarity to the desired atom type. For example, to identify the NHR3⁺ donor group, we evaluate the distance from the nitrogen (atom i in general) to all of the atoms in the vicinity. Each of these atoms contributes to the atom valence v_i by a factor determined by a function v_{ij} of the interatomic distance r_{ij} , which smoothly switches from 1 for covalent distance r_{cov} to 0 at $1.6 r_{\text{cov}}$. The factor 1.6 has been chosen here so that two atoms bound to the same center in a tetrahedral arrangement do not affect one other. The polynomial switching function described above (eq 3) is used:

$$v_{ij}(r_{ij}) = \begin{cases} 1 & \text{if } r_{ij} < r_{\text{cov}} \\ 0 & \text{if } r_{ij} > 1.6r_{\text{cov}} \\ f_{\text{sw}}((r_{ij} - r_{\text{cov}})/(0.6r_{\text{cov}})) & \text{otherwise} \end{cases} \quad (6)$$

$$v_i = \sum_{j \neq i} v_{ij}(r_{ij}) \quad (7)$$

The scaling for NHR3⁺ is applied if the valence of the nitrogen v_{N} is between 3 and 5; the scaling factor f_{charge} is linearly

dependent on the valence and peaks at the value of c_s determined for the minimum geometry when the valence reaches $v_{\text{max}} = 4$:

$$f_{\text{charge}} = \begin{cases} 1 + (c_s - 1) \times (1 - |v_{\text{N}} - v_{\text{max}}|) & \text{for } v_{\text{N}} \text{ in } (v_{\text{max}} - 1 \dots v_{\text{max}} + 1) \\ 1 & \text{otherwise} \end{cases} \quad (8)$$

In the COO[−] group, the scaling factor f_{charge} is the product of the contributions from both oxygens, calculated using the same functions as above but peaking at a valence of 1.0. Such a continuous description of the atom type can be applied to any other group.

Water-Donor Scaling. In PM6, interaction energies of hydrogen bonds with a water molecule as a donor (including the water dimer) are not as underestimated as in other hydrogen bonds featuring a –OH group donor. This is probably a consequence of the special focus on water in the parametrization of PM6. Therefore, a large improvement can be achieved when a water molecule as a donor is treated separately from other oxygen donors. We define the atom type for water oxygen as a continuous property analogously to the identification of the charged groups. The scaling factor f_{wat} is equal to an optimized coefficient c_{wat} where the oxygen has exactly two hydrogens within covalent distance and decays smoothly when the fractional valence diverges from the ideal value, as determined using eq 8.

DISPERSION CORRECTION

We adopt the formalism and atomic parameters proposed by Grimme for the DFT, the D3 dispersion correction.¹⁷ We do not use the higher-order contributions (effectively covered by the $1/r^8$ term in D3), as they do not improve the accuracy when used with the SQM methods. The dispersion is pairwise interatomic potential:

$$E_{\text{disp}} = s_6 \sum \sum_{i,j} \frac{c_{6,ij}}{r_{ij}^6} f_{\text{damp}}(r_{ij}) \quad (9)$$

where $f_{\text{damp}}(r_{ij})$ is a function damping the dispersion at short distances:

$$f_{\text{damp}}(r_{ij}) = \frac{1}{1 + 6 \left(\frac{r_{ij}}{s_r r_{0,ij}} \right)^{-\alpha}} \quad (10)$$

In DFT-D, it is used to prevent overbinding in the region where the exchange-correlation functional describes the interaction

well. In the SQM methods, the interaction at short range is also described by the method itself by means of parametrized core–core potentials.

There are three adjustable parameters that should be optimized for each corrected method—the global scaling of the correction s_6 , the scaling of the cutoff radii in the damping function s_r , and the exponent α that determines the steepness of the damping function.

The main difference from the previously used dispersion correction lies in the parameters employed in the formulas above, the c_6 coefficients that determine the dispersion energy for each pair of atoms and the cutoff radii r_0 determining the onset of the damping function that switches off the correction at short distances. Previously, the c_6 coefficients were tabulated for each element, and pairwise $c_{6,ij}$ coefficients were constructed using empirical combination rules. In D3, all of the pairwise coefficients have been calculated using time-dependent DFT not only for each pair of elements but also considering the possible different valence states of the atoms. To allow a change of the parameters during a reaction, the valence is defined as a continuous function of the coordinates of the surrounding atoms, and the $c_{6,ij}$ coefficient is interpolated from the values tabulated for the reference valences. Also the cutoff radii were determined by calculation as dispersion-specific pairwise radii, replacing the previously used sum of atomic van der Waals radii. For further details on the dispersion correction, we refer the reader to the original paper.

However, the dispersion correction itself does not yield satisfactory results when used with PM6 and other semiempirical methods. We found that there is a specific error in the description of hydrocarbons where the intermolecular distance is strongly underestimated owing to weak Pauli repulsion between hydrogens. This cannot be corrected by the dispersion, which is only attractive. Therefore, we had to add a repulsive term to all pairs of hydrogen atoms. We have chosen the form of a smooth sigmoidal function:

$$E_{\text{rep}}(r_{ij}) = s_{\text{HH}} \times \left(1.0 - \frac{1.0}{1.0 + \exp\left(-e_{\text{HH}}\left(\frac{r_{ij}}{r_{0,\text{HH}}} - 1.0\right)\right)} \right) \quad (11)$$

where s_{HH} sets the strength of the correction, $r_{0,\text{HH}}$ defines the distance where the function acts, and the exponent e_{HH} determines how steep it is. This function is flat at short distances and therefore does not affect the covalent-bond region in any other way than by adding a constant. At the close noncovalent region, it approximates the exponential repulsion well. While this repulsive correction is independent of the dispersion, in practical implementation it is calculated along with the dispersion correction and is included in the D3 term in our notation.

REFERENCE DATA

Training Set. Both corrections have been parametrized on the recently introduced S66x8 data set.²³ It features CCSD(T)/CBS dissociation curves for 66 noncovalent complexes, covering hydrogen bonds, dispersion, and mixed dispersion/electrostatic

interactions. The hydrogen bonds in the set cover all of the combinations of the most common donor/acceptor groups and include also cyclic double hydrogen bonds. Unlike the previously used S22 set,²⁴ it also provides a more balanced coverage of dispersion, including both π – π stacking and dispersion in aliphatic hydrocarbons. These improvements over the previously used benchmark data and availability of the dissociation curves instead of equilibrium geometries are a great advantage in this application and lead to a more robust resulting method.

Charged H-Bonds. To develop the H-bonding correction for systems with charged groups, we built a new data set consistent with the S66 set. It covers the charged moieties found in proteins, using model molecules acetate, methylammonium, guanidinium, and imidazolium (heterocycle in the protonated histidine), interacting with small donor/acceptor molecules (water, methanol, methylamine, and formaldehyde). Details on this set are provided in the Supporting Information. The geometries of the complexes and the benchmark CCSD(T)/CBS results are also available online in the BEGDB database²⁵ (www.begdb.com).

Validation Sets. We test the methods on multiple data sets covering noncovalent interactions in organic molecules and biomolecules. In addition to the S66x8 set used in parametrization, we report separately the results for the equilibrium geometries, the S66 set.²³ The same complexes are used in the S66a8 set²⁶ to cover those geometries distorted in the intermolecular angular coordinates. We also employ the S22 data set²⁴ (used to parametrize the previous generations of the corrections), as recalculated in the large basis set by Szalewicz.²⁷ Two separate sets are utilized for the validation of the H-bonding and dispersion separately: a set of 104 H-bonds (abbreviated as HB104 here) optimized and calculated at the MP2 level, developed for the parametrization of the original -DH correction,⁶ and a set of hydrocarbon dimers²⁸ (abbreviated as HC12). The latter covers at the CCSD(T)/CBS level the series propane to hexane, cyclopropane to cyclohexane, butadiene, and hexatriene and their cyclic analogs. Another test set, labeled here as AA24, is a set of neutral and charged amino acid side chains in the geometries most commonly found in proteins.²⁹

COMPUTATIONAL SETUP

The PM6, RM1, AM1, and PM3 calculations, including the DH+ correction, have been carried out in MOPAC 2009.³⁰ The OM3 method was used as implemented in the MNDO 2005 program.³¹ For the SCC-DFTB calculations, the DFTB+ program has been used.³²

The SCC-DFTB calculations use the third-order expansion and modified N–H parameters³³ that improve the interaction energies. In our comparison, we have also included the SCC-DFTB with the original dispersion correction³⁴ (denoted as -D) and with modified electrostatics of hydrogen (denoted as γ), which improves the hydrogen bonding.

The D3H4 correction energy is solely a function of the molecular geometry. All of the corrections here are independent of each other, and the corrected total energy is a sum of the energy from the semiempirical calculation (E_{SQM}) and the correction terms:

$$E_{\text{SQM-D3H4}} = E_{\text{SQM}} + E_{\text{HB}} + E_{\text{disp}} + E_{\text{rep}} \quad (12)$$

where E_{HB} is given by eq 1, E_{disp} by eq 9, and E_{rep} by eq 11.

The D3H4 correction developed here was implemented separately, so that it can be applied to results from any of the

programs above. This experimental code is based on the Cuby framework developed in our laboratory. We plan to implement these corrections in MOPAC.³⁰ A standalone program for calculation of the hydrogen bond correction is available at the author's Web site (www.molecular.cz/~rezac). A program implementing the D3 correction (without parameters for SQM methods) is available at the Web site of Grimme's group (toc.uni-muenster.de/DFTD3/).

PARAMETERIZATION

The dispersion correction is parametrized first on complexes without hydrogen bonds. Subsequently, the hydrogen-bonding correction is optimized, taking into account the contribution of dispersion to the H-bonded complexes.

All of the parametrizations described here are least-squares optimizations, minimizing the root-mean-square error of the interaction energy when compared to the CCSD(T)/CBS reference.

Dispersion. The parametrization of the dispersion correction is not trivial and cannot be fully automated. These problems are caused by the unbalanced description of different types of molecules given by the SQM methods and, in the case of PM6, also the partial coverage of the dispersion. Therefore, multiple separate steps are needed to get a robust set of parameters. The final balancing of the different types of interaction was done by hand. To obtain the parameters presented here, the following protocol was employed:

- (1) As we have already shown, in the case of PM6, it is necessary to introduce scaling of the dispersion correction energy. We derive the scaling coefficient s_6 from the long-distance interactions where the effect of the damping function is negligible. We have used the most distant points in the S66x8 data set (displaced to $2\times$ the equilibrium distance) of the dispersion group. The optimization of the coefficient yields a value of 0.88, which is almost the same value as that used in the previous generation of the correction. This scaling is not needed in the other methods investigated here (s_6 is 1.0).
- (2) In the next step, the damping function is optimized. For this, we use dispersion and mixed-type complexes from S66x8, excluding the aliphatic hydrocarbons that exhibit anomalous behavior, which is corrected later.
- (3) If no special measures are taken, the interaction between the aliphatic hydrocarbons is overestimated by all of the methods considered here. This effect is strongest in the case of PM6, the intermolecular distance becomes extremely short when optimized with not only PM6-D or PM6-D3 but also PM6 alone (1.5 Å hydrogen–hydrogen contact in the worst case). This problem cannot be fixed by tweaking the dispersion, but a separate repulsive correction has to be added. This repulsion has been optimized on the most problematic system in the S66 set, the neopentane dimer. The optimization of the function on other hydrocarbons yields too weak a repulsion to correct the geometry of the neopentane dimer. In order to introduce the least perturbation, we do not optimize all of the variables freely, but we seek the smallest s_{HH} for which one can obtain the correct shape of the dissociation curve while optimizing the remaining two parameters. The effect of the repulsive H–H correction on the dissociation curve of the hydrogen molecule

Table 3. The Dispersion Correction Parameters for the Methods Considered in This Study^a

parameter	PM6	SCC-DFTB	RM1	OM3	AM1	PM3
s_6	0.88	1.0	1.0	1.0	1.0	1.0
s_r	1.18	1.215	1.0	1.14	0.90	0.90
α	22	30	16	23	15	22
s_{HH} (kcal/mol)	0.4	0.3	0.3	0.3	0.9	0.9
e_{HH}	12.7	14.31	4.46	9.60	4.46	6.86
$r_{0,\text{HH}}$ (Å)	2.30	2.35	2.11	2.10	2.11	2.23

^aThe parameters are dimensionless unless indicated otherwise.

dimer is illustrated in plot S1 in the Supporting Information. Even in this simple system, this correction is necessary for reproducing both the interaction energy and intermolecular distance of the equilibrium structure.

- (4) This repulsive correction now leads to an underestimation of the interaction in all of the hydrocarbons except for neopentane. Here, no universal solution that works for all systems can be found. Therefore, we have manually adjusted the strength of the repulsive term (by changing the s_{HH} parameter) to get the best interaction energies overall with the condition of conserving a reasonable (within 5% from the benchmark) intermolecular distance in the neopentane dimer.

The final set of parameters is listed in Table 3. The final performance of this correction is negligibly worse (by 0.01 kcal/mol in the S66 dispersion complexes in PM6) than the solution obtained by a blind optimization of the dispersion correction without the repulsive term, but the description of the hydrocarbons is improved significantly, bringing more balanced errors in the different types of interactions.

Hydrogen Bonding. The first step of the development of the H-bonding correction was the design of the functional form, mainly the shape of the radial and angular terms. Here, we have attempted to build functions with the desired properties described above, matching the distance and angular dependence of the error between PM6 and the reference data. For the radial term, we have used the dissociation curves of the hydrogen-bonded complexes in the S66x8 data set; the angular term was optimized on angular scans in methanol and methylamine dimers.

Once the form of the potential is set, it is straightforward to optimize the free parameters, the coefficients determining the strength of the correction. The coefficients for all donor–acceptor combinations (c_{OO} , c_{ON} , c_{NO} , c_{NN}) have been optimized along with the coefficient for scaling the H-bonds with the water donor, c_{wat} , on the hydrogen-bonded complexes in the S66x8 data set.

Finally, the scaling coefficients for the charged groups present in the training set (carboxylic acids, ammonium, guanidinium and imidazolium), $c_{\text{s,COO-}}$, $c_{\text{s,NHR3+}}$, $c_{\text{s,gua}}$ and $c_{\text{s,imz}}$, have been optimized on a set of charged hydrogen bonds. The resulting parameters are listed in Table 4. Note that the parameters for different cations differ significantly; it is not possible to use a single parameter and achieve the desired accuracy here.

We have attempted to reoptimize some of the parameters in the functional form of the correction, e.g., the radii defining the radial potential, on the S66x8 set. Although the overall error can be slightly decreased this way, the geometric parameters are worse (the minima become shifted away from the reference geometry). Therefore, we keep the original radial function

designed to have the optimal shape to ensure a robust description of the complex geometries.

RESULTS AND DISCUSSION

Tests on Benchmark Data. The corrected SQM methods developed here have been tested on multiple benchmark data sets. Table 5 lists the RMSE for all of the tested methods and sets. Other error measures, the mean and maximum unsigned errors, are listed in Tables S2 and S3 in the Supporting Information. This overview includes sets used for the parametrization of the corrections in some of the methods (S66x8 and the charged H-bonds for D3H4 and S22 for the DH2 and DH+ corrections).

Table 4. The Hydrogen-Bonding Parameters for the Methods Considered in This Study^a

parameter	PM6	SCC-DFTB	RM1	OM3	AM1	PM3
c_{OO} (kcal/mol)	2.32	1.11	3.76	1.95	4.89	2.71
c_{ON} (kcal/mol)	3.10	2.58	3.90	1.64	6.23	4.37
c_{NO} (kcal/mol)	1.07	0.80	3.14	0.93	2.54	2.29
c_{NN} (kcal/mol)	2.01	2.01	2.95	1.35	4.56	3.86
c_{wat}	0.42	1.32	0.94	0.50	0.49	0.91
$c_{\text{s,COO-}}$	1.41	1.22	1.10	1.63	1.08	0.89
$c_{\text{s,NHR3+}}$	3.61	2.33	1.21	0.9	2.78	2.54
$c_{\text{s,gua}}$	1.26	2.42	1.18	1.37	0.86	1.54
$c_{\text{s,imz}}$	2.29	3.44	1.10	1.18	2.11	1.84

^a The parameters are dimensionless unless indicated otherwise.

The average of the errors over all of the data sets (E_{all}) is used as a simple measure of the overall performance of each method. Additionally, the average over the sets that were not used in the parametrization of any method (HB104, large set of H-bonds; hydrocarbon dimers, HC12 and AA24, amino acid side chains), E_{val} , is provided as an independent validation. The optimization of the new correction for the hydrogen bonds of the charged groups leads to an important decrease of error in these cases. Here, a separate parametrization is needed for each functional group. The data set listed here is the one used for parametrization and covers all of the charged amino acids in proteins.

When different corrections are compared for each SQM method, the D3H4 approach developed here yields the best results both on the validation sets and overall. In the following text, we have ordered the methods by their overall best score, discussing the results in detail and comparing the possible correction schemes for each given method. Here, we have also discussed the advantages and disadvantages of each method for practical applications.

Among the methods tested, OM3-D3H4 yields the lowest errors (E_{all} 0.69 kcal/mol, E_{val} 0.63 kcal/mol), which is an improvement of about 35% over OM3-DH2. We cannot compare OM3-DH+ here, because we do not have software that implements this combination, but on the basis of the original paper, we expect it to be somewhere between -DH2 and -D3H4. Although OM3-D3H4 scores best for interaction energies, there are two issues that make it impractical for many applications. First, the OM3 method is parametrized for only a few elements

Table 5. The Root Mean Square Errors (in kcal/mol) of the Studied Methods in Multiple Benchmark Data Sets^a

	S66	S66x8	S66a8	S22	H bonds	charged HB	hydrocarbons	aaside chains	avg	avg _{testing}
PM6	3.02	2.49	2.12	4.16	3.18	3.92	2.64	4.08	3.20	3.30
PM6-DH2	0.91	0.79	0.73	0.54	1.52	2.21	0.67	1.32	1.09	1.17
PM6-DH+	0.82	0.76	0.67	0.80	1.43	1.94	0.67	1.89	1.12	1.33
PM6-D3H4	0.65	0.66	0.68	0.78	1.05	1.11	0.71	1.17	0.85	0.98
PM6-D3H4*	0.70	0.71	0.74	0.84	1.12	2.26	0.71	1.86	1.12	1.23
DFTB	2.88	2.40	2.24	3.45	2.82	4.78	2.90	3.44	3.11	3.05
DFTB-D	1.50	1.43	1.28	1.63	1.96	4.28	0.59	2.27	1.87	1.60
DFTB-D, γ	1.17	1.17	1.04	1.21	1.61	3.67	0.56	1.82	1.53	1.33
DFTB-DH2	1.44	1.15	0.98	1.86	1.54	2.13	0.59	1.62	1.41	1.25
DFTB-D3H4	0.67	0.62	0.61	0.97	0.71	1.43	0.59	0.88	0.81	0.73
RM1	5.39	4.38	4.13	7.15	5.40	5.60	3.65	5.34	5.13	4.80
RM1-D3H4	0.92	0.90	0.78	1.03	0.90	2.05	0.24	0.73	0.94	0.62
RM1-D3H4*	0.91	0.90	0.79	1.03	0.89	2.09	0.24	0.93	0.97	0.69
OM3 ^b	3.33	2.70	2.49	4.17	2.88	3.00	3.93	4.99	3.44	3.93
OM3-DH2 ^b	0.80	0.96	0.62	0.96	0.84	1.83	1.11	1.53	1.08	1.16
OM3-D3H4 ^b	0.48	0.60	0.42	0.58	0.56	1.50	0.70	2.34	0.90	1.20
AM1	6.24	5.27	4.03	8.66	6.10	7.64	3.73	6.38	6.01	5.40
AM1-DH2	1.93	1.96	1.47	0.85	2.08	3.58	3.94	3.71	2.44	3.25
AM1-D3H4	1.35	1.76	1.45	1.76	2.11	3.04	0.82	2.02	1.79	1.65
PM3	5.08	4.51	3.77	7.64	4.98	7.03	2.25	4.60	4.98	3.94
PM3-D3H4	1.40	1.26	0.97	2.51	0.83	2.23	0.40	1.05	1.33	0.76
B3LYP/6-31G*	2.68	2.40	1.87	3.63	1.31	3.17	4.20	2.97	2.78	2.82
TPSS/TZVP-D	0.69	0.53	0.57	0.58	1.04	1.89	0.72	0.89	0.86	0.88
BLYP/def2TZVP-D3	0.25	0.17	0.21	0.33	0.41	0.59	0.21	0.46	0.33	0.36
MP2/cc-pVTZ	0.70	0.59	0.57	1.85	1.40	1.81	0.88	1.62	1.18	1.30

^a The last two columns list the average of these errors over all of the sets and over the validation sets only. ^b Due to the limited parameter set, the methionine complexes in AA side chains set are not considered.

(H, C, N, and O). More importantly, this method critically fails in geometry optimizations of complexes containing acetic acid, as described in the following section.

The second most accurate method (E_{all} 0.81 kcal/mol, E_{val} 0.73 kcal/mol) is SCC-DFTB when used with the D3H4 correction and modified parameters for hydrogen–nitrogen interaction.³³ This is a clear improvement over DH2 but also over the original dispersion correction³⁴ and modified electrostatics (SCC-DFTB-D, γ) developed by the authors of DFTB to improve the description of the hydrogen bonds.³³

PM6-D3H4 scores third overall (E_{all} 0.85 kcal/mol, E_{val} 0.98 kcal/mol). The parametrization of PM6-D3H4 on the S66x8 set leads to very good results for all of the sets in the S66 family. Achieving such low errors with a substantially simplified H-bonding correction is a very encouraging result indicating a good choice of the form of the correction. The performance on the S22 set is better than that of PM6-DH+ but not as good as that of PM6-DH2 (both of these methods used S22 for parametrization). What is more important is the independent tests on systems outside the training set. In the set of 104 H-bonds, PM6-D3H4 outperforms all of their predecessors with a RMSE of 1.1 kcal/mol. This is the most important result, which clearly shows the accuracy and robustness of the new H-bonding correction. In the set of hydrocarbon dimers, the results are slightly worse than in the previous version, but the D3 correction corrects the problems with short intermolecular distances in aliphatic hydrocarbon dimers. Although PM6 has some limitations, the accuracy that can be achieved for noncovalent interactions, in combination with the coverage of a major part of the periodic table (both by PM6 and the D3 dispersion), makes PM6-D3H4 very useful for applications.

To demonstrate the effects of the introduction of the water atom type and the scaling of charged hydrogen bonds, we list results obtained without these modifications (PM6-D3H4* in Table 5). In the case of water as a hydrogen bond donor (included in S66, S22, and H-bonds sets), the increase of the overall errors is rather small, but the error for the water dimer is rather large (the binding is overestimated by 1.2 kcal/mol, which is 25% of the interaction energy). We are convinced that hydrogen bonds in and with water are important in many applications, and correcting this error is worth the specific scaling in the H-bond correction. Regarding the hydrogen bonds in charged systems, the error is about twice as large as when the scaling is applied. A RMSE of 2.26 kcal/mol translates to rather small relative errors, as the interaction energies in these systems are larger than in neutral H-bonds. The improvement brought by the system-specific scaling allowed us to achieve high accuracy consistently in a wide range of systems. When the ultimate accuracy is not needed, it is possible to use the H-bond correction without this scaling; in such a case, the errors are comparable to the previous generations of the correction.

In this paper, the RM1 method has been coupled with empirical corrections for the first time, and the results are very promising. The error in the validation sets is very low (0.62 kcal/mol), but the overall results are slightly worse because of the large error in the charged H-bonds. Unlike all of the other methods, RM1 works comparably well without separate scaling of the H-bonds with the water donor and in charged H-bonds (the method is denoted as RM1-D3H4*). This system-independent nature of the errors indicates that the method is robust and not overparameterized. We plan to test this method in applications where the limited set of parameters (H, C, N, O, P, S, and halogens) makes it possible.

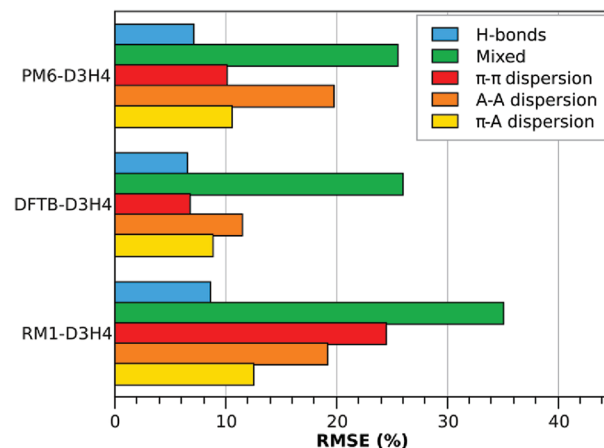


Figure 3. The relative errors of the selected methods for interactions of different types in the S66 set: H-bonds; mixed type; and π - π , aliphatic–aliphatic, and π -aliphatic dispersion. Errors are plotted as a percent of the average interaction energy in the group.

The AM 1 and PM3 methods were included only for comparison; it is obvious from the results that the more modern SQM methods can describe noncovalent interactions better. Here, PM3 performs better than AM1 but still yields rather large errors for some data sets.

Table 5 also lists results of selected DFT and wave function methods for comparison. The widely used B3LYP/6-31G* setup was chosen as a representative DFT method without any treatment of dispersion. DFT with empirical dispersion is represented by the TPSS/TZVP-D method,¹⁶ which yields very good results with a reasonably small basis set, and BLYP/def2-QZVP with the D3 dispersion correction¹⁷ using the Becke–Johnson damping function³⁵ as a demonstration of the highest accuracy that can be achieved with DFT-D when large basis set and advanced dispersion corrections are used. Finally, MP2 with the cc-pVTZ basis set (with counterpoise correction of basis set superposition error) was chosen as an example of a relatively inexpensive correlated QM method. The best corrected semiempirical methods (PM6-D3H4, DFTB-D3H4, and OM3-D3H4) outperform DFT and MP2 and are about as accurate as DFT-D with a medium-sized basis set.

For the selected method that performed best, we have analyzed the errors in the S66 data set in more detail. The set is divided into five groups—H-bonds; π - π , π -aliphatic, and aliphatic–aliphatic dispersion; and mixed-type interactions. For each group, the relative error is calculated as RMSE expressed in the percentage of the average interaction energy in the group to make the errors comparable between interactions of different strengths. The results are plotted in Figure 3. The hydrogen bonds are described very well by all of these methods. In DFTB-D3H4 and PM6-D3H4, the largest error in the dispersion complexes is found in the complexes of aliphatic hydrocarbons, because these complexes exhibit very large errors already in the uncorrected methods. RM1-D3H4 has the largest error among the dispersion groups for π - π interactions, where the stacking interactions of uracil are underestimated by as much as 1.7 kcal/mol. It is not surprising that the largest errors are observed in the mixed-type groups, where the interactions are not specifically corrected (e.g., X- π interactions and C–H hydrogen bonds).

Geometries. Another important test of the corrected method is its application to the optimization geometry of noncovalent

Table 6. The Results of the Geometry Optimization of the S66 Complexes by the Studied Method^a

	avg. RMSD	max. RMSD	RMSE S66	RMSE S66 _{opt}
PM6	0.38	2.34	3.1	2.7
PM6-D/PM6-DH2	0.26	0.85	0.9	1.0
PM6-DH2 (numer.)	0.22	0.73	0.9	1.3
PM6-DH+	0.25	1.30	0.8	1.0
PM6-D3/PM6-D3H4	0.25	1.30	0.7	0.7
PM6-D3H4	0.20	0.51	0.7	0.7
DFTB	0.37	1.79	2.9	2.3
DFTB-D	0.24	0.83	1.6	1.1
DFTB-D, γ	0.23	0.77	1.2	0.8
DFTB-D3H4	0.21	0.77	0.7	0.9
RM1	0.55	5.08	5.4	4.4
RM1-D3H4	0.23	1.01	1.0	0.9
RM1-D3H4*	0.23	0.93	0.9	0.9
OM3	0.65	6.36	3.4	5.9
OM3-D/OM3-DH2	0.24	1.02	0.9	6.3
OM3-D3H4	0.20	0.73	0.5	5.0
AM1	0.80	5.95	6.3	4.1
AM1-D3H4	0.39	2.23	1.4	2.0
PM3	0.58	2.23	5.1	4.0
PM3-D3H4	0.37	2.26	1.4	1.2
TPSS/TZVP-D	0.11	1.35	0.7	1.3

^aWe describe the changes in geometry by the average and largest root-mean-square deviation (RMSD, in Å) and the changes in the interaction energy by listing the RMSE (in kcal/mol) in the S66 set before and after geometry optimization.

complexes. We evaluate two criteria: First, the optimized geometry should be as close as possible to a benchmark one optimized with a high-level QM method. Second, the interaction energies calculated on the optimized geometries should be close to the reference values calculated at reference energy. When these two criteria are satisfied, the method is applicable to practical calculations that involve geometry optimization and an evaluation of the properties of the resulting structure.

We performed these tests on the S66 data set. We optimized each of the complexes with the studied method with high accuracy (convergence limits of 0.03 kcal/mol/Å for the RMS gradient, 0.06 kcal/mol/Å for the max. gradient component, and a 3.0e−4 kcal/mol energy difference between the subsequent steps). The interaction energies are recalculated on the new geometries and compared to the benchmark. The results are summarized in Table 6. We list the average and largest root-mean-square deviations (RMSD, in Ångstrom) compared to the reference MP2/cc-pVTZ (counterpoise corrected) geometry along with the RMSE of the interaction error on the benchmark geometries and after optimization with the tested methods.

In some cases, the optimization is problematic. The DH2 correction uses only an approximate gradient, and it is not possible to converge the optimizations. Instead, we list the results of the optimizations with dispersion only, but we calculate the interaction energies with both dispersion and H-bond correction; this is the protocol we have used and recommend for applications of this method. For PM6-DH2, we also performed full optimization using the much more expensive numerical evaluation of the gradient. The DH+ correction fixes this issue, but the gradient is not smooth. The cusp corresponds to the

Table 7. The Interaction Energy (in kcal/mol) Per Molecule in a Cubic Box with 216 Water Molecules^a

	ΔE /molecule	ΔE_{dimer}
PM6	−5.2	−3.9
PM6-DH2 angle only	−9.5	−4.9
PM6-DH2	−8.4	−4.9
PM6-DH+	−9.6	−6.5
PM6-D3H4	−8.3	−4.9
TIP3P	−7.4	−6.0
CCSD(T)		−5.0

^aThe interaction energy in the water dimer is listed for comparison.

linear arrangement; therefore, it is impossible to converge the optimizations of symmetric systems (namely, complexes 1, 8, 17, and 23). Since the resulting structure is very close to the minimum although the gradient is nonzero, we use the unconverged geometries for further analysis.

A serious issue is observed in the OM3 method, regardless of whether the corrections are applied or not. The complexes containing acetic acid optimize into covalently fused structures where the hydrogen has an equal distance of 1.2 Å from the donor and acceptor atoms. The interaction energy in these cases is exaggerated by 100 and 200%. If this issue were removed, the method would perform rather well, as indicated by the low average RMSD.

The D3H4 consistently yields the lowest average and maximum RMSD when compared to the other possible correction schemes. The improvement in PM6, where PM6-D3H4 yields very low maximum RMSE compared to its predecessors, is an important achievement. This is partly due to the special treatment of the dispersion in aliphatic hydrocarbons, where the newly introduced repulsive correction is needed to obtain good geometries. Also, the new H-bonding correction works slightly better, although it is simpler than the previous versions and uses less information from the local geometry of the hydrogen bond (see Table 1). The average RMSD in the 23 H-bonds in the S66 set is 0.24 Å for PM6-DH+ and improves only slightly to 0.23 Å in PM6-D3H4, but the largest RMSD decreases from 0.94 to 0.51 Å. What is also very important is that the interaction energies do not change significantly when the structures are optimized.

We applied the same optimization protocol to DFT-D¹⁶ in a medium-sized basis set (TZVP). This method is comparable to the corrected semiempirical methods in terms of interaction energies. The results of geometry optimizations are better overall (the average RMSD is 0.11 Å while the best SQM-D3H4 methods yield 0.2 Å). A slightly larger maximum RMSD and an increase of the error in interaction energies in the optimized complexes is caused by a single system, where the methylamine–methanol complex with amine hydrogen bond donor optimizes to the global minimum with the alcohol as a hydrogen donor.

Tests on Large Systems. As the corrections are developed on small model complexes, it is necessary to evaluate their transferability to large systems. The dispersion correction should have no problems here, as the dispersion interaction is almost additive and the pairwise potential used is a good approximation to it. The scaling of the whole correction is obtained from calculations of the complexes displaced to twice the equilibrium geometries in order to ensure that the dispersion is not overestimated at longer distances.

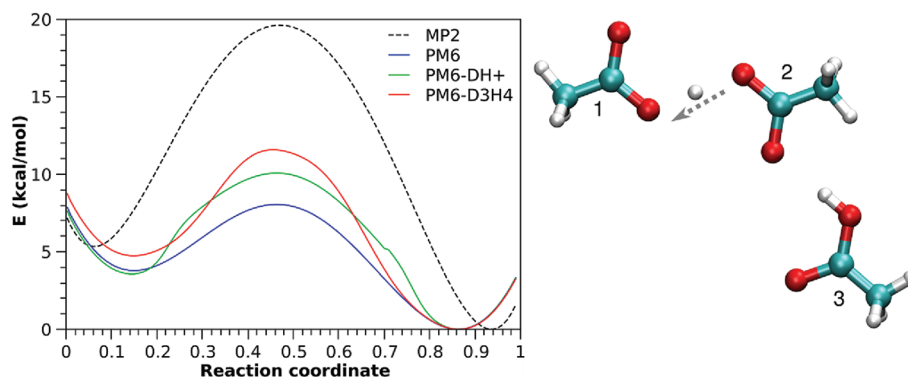


Figure 4. The proton transfer in a network of hydrogen bonds. The energy profile along the reaction coordinate obtained with PM6 (blue), PM6-DH+ (green), and PM6-D3H4 (red) is compared with the MP2/aug-cc-pVDZ reference (black).

More attention has to be paid to the hydrogen bonds. In condensed systems, the number of potential H-bonds grows rapidly, and even if they contribute only negligibly to the total energy, the overall stabilization arising from the H-bond correction might be overestimated. This problem, observed in the first generation of correction, was addressed by a more complex definition of the hydrogen bonds in DH2 and DH+, using additional internal coordinates. Here, we have attempted to solve this problem by making the correction rather short-ranged, which eliminates all of the false contributions from potential H-bonds other than those actually interacting.

We test this on a system with very high H-bond density—water. We calculate the total interaction energy in a cubic box of 216 water molecules and list it as interaction energy per single water molecule. In Table 7, we compare the results of PM6 with the DH2, DH+, and D3H4 corrections with PM6-DH2 without the additional angular and torsional coordinates (using the H-bond angle α only). The interaction energies in the optimized structure of a water dimer are included for comparison. For reference, we list the values obtained with the TIP3P force field and accurate CCSD(T) interaction energy.

These results show that the new approach is as efficient as the use of additional coordinates in the DH2 correction, while the new correction is much simpler and can be calculated more efficiently. In this case, PM6-DH+ yields larger average interaction in the cluster, although it uses a formalism very similar to DH2. This is caused by the overestimated interaction of the H-bond of this type already in the water dimer.

Proton Transfer. The new hydrogen-bonding correction can seamlessly describe proton transfer, because it smoothly switches the donor and acceptor when the hydrogen atom is in the middle. The description is more complicated when the proton transfer studied involves a charged group, because the scaling of the H-bonding correction applied to that group changes as well. Using the continuous scaling introduced above, a smooth potential is obtained even in the most complex cases. This is illustrated in Figure 4 on the potential energy curve along a proton transfer in a network of carboxylic groups with a total charge of -1 , with the MP2/aug-cc-pVDZ calculation serving as a reference. In this simple model, all of the coordinates have been fixed, while the proton is transferred along the H-bond axis. In this system, central molecule 2 becomes charged as it loses the proton, which makes the second hydrogen bond between molecules 2 and 3 stronger. The correction stabilizes the minima but does not affect the energy of the transition state, which effectively increases the barrier. This is an

improvement toward the reference curve when compared to uncorrected PM6, although the barrier height is still underestimated. We also include the PM6-DH+ results for comparison. Unlike its predecessors, this method should yield a smooth potential energy curve. In the practical implementation in MOPAC2009, there is a minor discontinuity close to a reaction coordinate value of 0.7. This most probably arises from the use of a distance cutoff in the H-bond correction. Overall, the shape of the curve, the energy difference between the minima, and the barrier height are not as good as in PM6-D3H4.

CONCLUSIONS

Empirical corrections for noncovalent interactions can substantially improve the performance of semiempirical quantum mechanical methods, reaching chemical accuracy (error of 1 kcal/mol) in most of the benchmark data sets studied. These results are very close to much more expensive methods, such as DFT-D or MP2, while the efficiency of the SQM method makes it possible to study very large systems on a routine basis.

The accuracy of the corrected SQM method approached its limits already with the DH2 correction, and the later advancements including the one presented here aim mainly to improve the robustness of the method. This was achieved by adopting the latest developments in the dispersion corrections for the DFT methods and redesigning the H-bonding correction from scratch. Although the H-bonding correction has been substantially simplified, the accuracy was improved.

We have addressed multiple weaknesses of the previous generations of the corrections. Most importantly, the D3H4 correction is the first one that can be used for geometry optimizations and molecular dynamics, as it and its derivatives have a continuous and smooth potential energy surface. For the first time, we have used scaling of the correction in charged hydrogen bonds in order to improve the accuracy in these systems.

The new H-bond correction does naturally describe proton transfer along a hydrogen bond, yielding a smooth potential energy surface even in the most complex cases. Stabilization of the minima effectively increases the barrier height, improving the SQM results toward a more accurate reference.

Among the tested methods, PM6-D3H4, DFTB-D3H4, and RM1-D3H4 yield errors lower than 1 kcal/mol in multiple benchmark data sets. We have also shown that these methods reproduce geometries of noncovalent complexes with good accuracy, which makes them useful for many applications.

Semiempirical methods with corrections for noncovalent interaction can yield very accurate results on small model systems and have been successfully applied to real-world systems. However, we would like to end this paper with a warning: The accuracy of both the SQM methods and of the corrections is achieved by empirical parametrization, and they can yield large errors when applied to systems that are outside of this parametrization. Therefore, it is advised to examine the results critically and possibly check them against more reliable calculations.

■ ASSOCIATED CONTENT

S Supporting Information. Description and geometries of the charged hydrogen bonds data set and additional tables listing mean and maximum errors for all the studied methods in multiple benchmark data sets are provided as Supporting Information. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*Fax: +420 220 410 320. E-mail: rezac@uochb.cas.cz.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENT

This work was a part of Research Project No. Z40550506 of the Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, and was supported by Grant No. MSM6198959216 from the Ministry of Education, Youth and Sports of the Czech Republic. It was also supported by the Research and Development for Innovations Operational Program of the European Social Fund (CZ.1.05/2.1.00/03.0058). The support of Praemium Academiae, Academy of Sciences of the Czech Republic, awarded to P.H. in 2007 is also acknowledged. We are grateful to James Stewart for sharing his knowledge of semiempirical methods, to Martin Korth for discussion on testing of the H-bond correction in condensed systems, and to Filip Lankaš for his help with the construction of the polynomial functions.

■ REFERENCES

- (1) Stewart, J. J. P. *J. Mol. Model.* **2007**, *13*, 1173–1213.
- (2) Rocha, G. B.; Freire, R. O.; Simas, A. M.; Stewart, J. J. P. *J. Comput. Chem.* **2006**, *27*, 1101–1111.
- (3) Tuttle, T.; Thiel, W. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2159–2166.
- (4) Stewart, J. J. P. *J. Mol. Model.* **2008**, *15*, 765–805.
- (5) McNamara, J. P.; Hillier, I. H. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2362.
- (6) Řezáč, J.; Fanfrlík, J.; Salahub, D.; Hobza, P. *J. Chem. Theory Comput.* **2009**, *5*, 1749–1760.
- (7) Korth, M.; Pitoňák, M.; Řezáč, J.; Hobza, P. *J. Chem. Theory Comput.* **2010**, *6*, 344–352.
- (8) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.
- (9) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260.
- (10) Korth, M. *J. Chem. Theory Comput.* **2010**, *6*, 3808–3816.
- (11) Fanfrlík, J.; Bronowska, A.; Řezáč, J.; Přenosil, O.; Konvalinka, J.; Hobza, P. *J. Phys. Chem. B* **2010**, *114*, 12666–12678.
- (12) Pecina, A.; Přenosil, O.; Fanfrlík, J.; Řezáč, J.; Granatier, J.; Hobza, P.; Lepšík, M. *Collect. Czech. Chem. Commun.* **2011**, *76*, 457–479.
- (13) Dobeš, P.; Fanfrlík, J.; Řezáč, J.; Otyepka, M.; Hobza, P. *J. Comput.-Aided Mol. Des.* **2011**.
- (14) Dobeš, P.; Řezáč, J.; Fanfrlík, J.; Otyepka, M.; Hobza, P. *J. Phys. Chem. B* **2011**, *115*, 8581–8589.
- (15) Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1463–1473.
- (16) Jurečka, P.; Černý, J.; Hobza, P.; Salahub, D. *J. Comput. Chem.* **2007**, *28*, 555–569.
- (17) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *J. Chem. Phys.* **2010**, *132*, 154104.
- (18) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 221–264.
- (19) Clark, T. *THEOCHEM* **2000**, *530*, 1–10.
- (20) Winget, P.; Selcuki, C.; Horn, A. H. C.; Martin, B.; Clark, T. *Theor. Chem. Acc.* **2003**, *110*, 254–266.
- (21) Jug, K.; Geudtner, G. *J. Comput. Chem.* **1993**, *14*, 639–646.
- (22) Zhang, P.; Fiedler, L.; Leverentz, H. R.; Truhlar, D. G.; Gao, J. *J. Chem. Theory Comput.* **2011**, *7*, 857–867.
- (23) Řezáč, J.; Riley, K. E.; Hobza, P. *J. Chem. Theory Comput.* **2011**, *7*, 2427–2438.
- (24) Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985.
- (25) Řezáč, J.; Jurečka, P.; Riley, K. E.; Černý, J.; Valdes, H.; Pluháčková, K.; Berka, K.; Řezáč, T.; Pitoňák, M.; Vondrášek, J.; Hobza, P. *Collect. Czech. Chem. Commun.* **2008**, *73*, 1261–1270.
- (26) Řezáč, J.; Riley, K. E.; Hobza, P. *J. Chem. Theory Comput.* **2011**, *7*, 3466–3470.
- (27) Podeszwa, R.; Patkowski, K.; Szalewicz, K. *Phys. Chem. Chem. Phys.* **2010**, *12*, 5974.
- (28) Granatier, J.; Pitoňák, M.; Hobza, P. Unpublished data.
- (29) Berka, K.; Laskowski, R.; Riley, K. E.; Hobza, P.; Vondrášek, J. *J. Chem. Theory Comput.* **2009**, *5*, 982–992.
- (30) Stewart, J. J. P. *MOPAC 2009*; Stewart Computational Chemistry: Colorado Springs, CO, 2009.
- (31) Thiel, W. *MNDO 2005*; Max Planck Institute for Coal Research: Mülheim, Germany, 2005.
- (32) Aradi, B.; Hourahine, B.; Frauenheim, T. *J. Phys. Chem. A* **2007**, *111*, 5678–5684.
- (33) Yang, Yu, H.; York, D.; Cui, Q.; Elstner, M. *J. Phys. Chem. A* **2007**, *111*, 10861–10873.
- (34) Elstner, M.; Hobza, P.; Frauenheim, T.; Suhai, S.; Kaxiras, E. *J. Chem. Phys.* **2001**, *114*, 5149.
- (35) Grimme, S.; Ehrlich, S.; Goerigk, L. *J. Comput. Chem.* **2011**, *32*, 1456–1465.

Alternative Mechanisms in Hydrogen Production by Aluminum Anion Clusters

Paul N. Day,^{*,†,‡} Kiet A. Nguyen,^{†,§} and Ruth Pachter^{*,†}

[†]Materials and Manufacturing Directorate, Air Force Research Laboratory, Wright Patterson Air Force Base, Ohio 45433, United States

[‡]General Dynamics Information Technology, Inc., Dayton, Ohio 45431, United States

[§]UES, Inc., Dayton, Ohio 45432, United States

 Supporting Information

ABSTRACT: Possible mechanisms for the reaction of aluminum anion clusters with water have been studied theoretically using density functional theory for four different size clusters. Our results confirm the previously found (Reber et al. *J. Phys. Chem. A* **2010**, *114*, 6071) importance of Lewis-acid and Lewis-base sites on the cluster in the size specificity of the reactivity. However, alternative viable mechanisms have been found using both Langmuir–Hinshelwood and Eley–Rideal kinetics. Grotthuss-like mechanisms appear to be the most energetically favorable. We show that while the superatom theory successfully predicts reactivity of smaller clusters, it is less useful for the larger clusters.

I. INTRODUCTION

The reaction of aluminum with water to produce hydrogen gas is of interest as an alternative energy source. While aluminum in bulk reacts too slowly for practical applications, evidence suggests much faster rates for micro- or nano-sized aluminum particles.^{1,2} As with other metal nanoparticles, structure and properties vary by cluster size, and these variations may be at least partially explained or predicted through the use of “super-atom” theory and “magic numbers”, which are related to the spherical jellium model.^{3–8} In the superatom theory, the valence electrons in a cluster of metal atoms are sufficiently delocalized such that the wave function solution, in analogy with atomic wave functions, fills “super-atom” electronic shells designated as 1S, 1P, 1D, 2S, 1F, 2P, 1G, 2D, 3S, ..., thus generating the following series of magic numbers: 2, 8, 18, 20, 34, 40, 58, 68, 70, ... A cluster with a magic number of valence electrons should be particularly stable, in analogy with an inert gas. The aluminum atom has three valence electrons, and the superatom theory correctly predicts the inertness of Al_{13}^{-1} , with 40 valence electrons, as well as of Al_{11}^{-1} , with 34 valence electrons. The high reactivity of Al_{12}^{-1} , with 37 valence electrons, and of Al_{17}^{-1} , with 52 valence electrons, is also consistent with the theory.

The production of hydrogen gas from the reaction of aluminum nanoclusters with water was observed in a fast-flow reactor.^{9,10} Reber et al.¹⁰ investigated possible correlations between the reactivity with water and various calculated properties of the clusters, including dipole moment, binding energy, transition state energy, product energy, and orbital energies. The Al_{12}^{-1} cluster has a relatively large dipole moment and reacts rapidly with water (although apparently does not produce H_2), but the symmetric Al_{17}^{-1} has a zero dipole moment and also reacts rapidly with water, including the release of H_2 .

The Al_{12}^{-1} cluster also has a large binding energy with water and a low barrier for the OH bond-breaking, but these properties alone are not sufficient to predict the reactivity for each cluster size. The energy and structure of the Kohn–Sham molecular

orbitals were found to be important, particularly that of the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO). In systems with an odd number of electrons (Al_{12}^{-1} and Al_{20}^{-1}), the singularly occupied molecular orbital will be labeled SOMO, the lowest completely unoccupied orbital LUMO, and the highest doubly occupied orbital SOMO–1. Positions on the cluster where the LUMO protrudes out in space are Lewis-acid sites, which tend to attract the lone-pair of electrons of the oxygen atom in the water molecule. A water molecule will bind at these sites with a typical stabilization energy of 0.3–0.6 eV. If an adjacent aluminum atom has a strong contribution from the HOMO (the SOMO for odd-electron species), it can act as a Lewis-base, and one of the hydrogen atoms from the water molecule can bind to it, breaking its bond to the oxygen atom, resulting in the H and the OH being bound on adjacent aluminum atoms. For some clusters, the barrier for this reaction is less than the stabilization energy of the initial water binding, and this step is exothermic by over 1.0 eV, making the reaction thermally favorable. Other water molecules can react with other Lewis-acid–Lewis-base pairs on the aluminum cluster. Because of the exothermicity of this reaction, the system may have enough energy for two hydrogen atoms on adjacent aluminum atoms to form a bond and be released as H_2 . This is the Langmuir–Hinshelwood (LH) mechanism described by Reber et al.¹⁰ Alternative mechanisms that they describe include the Eley–Rideal (ER) type, where the second water molecule does not undergo bond-breaking on the surface but instead transfers a hydrogen atom directly to the bound hydrogen to form H_2 , and a “direct” mechanism, where neither water molecule undergoes surface bond-breaking but instead each directly contributes a hydrogen atom to form H_2 . In studies utilizing molecular dynamics (MD),^{11,12} a lower barrier for the first step of the reaction was found through a Grotthuss-like

Received: September 27, 2011

Published: November 22, 2011

mechanism where one, two, or three extra water molecules assist in the transfer of the hydrogen atom. This mechanism may be combined with either the LH or the ER mechanism to form a complete path for the formation of H₂.

In this study, we investigate the reaction of water with aluminum cluster anions of size $n = 12, 17, 20,$ and 23 , by using density functional theory and searching for viable reaction paths to produce H₂, taking all of the possible mechanistic paths into account. The reactions with $n = 12$ and $n = 17$ have been studied previously both experimentally and theoretically, and we expand upon those results. For $n = 23$, the number of valence electrons corresponds to the magic number of 70, so the superatom theory predicts it should be inert, but experiment has reported it to be highly reactive with water.¹⁰ Conversely, Al₂₀⁻¹, with 60 valence electrons, should be reactive, but has been reported to have low reactivity.¹⁰ Apparently, other structural or dynamic factors prevail over the superatom theory for these clusters. This is consistent with the results of Ma et al.,¹³ where in a study of the photoelectron spectra of aluminum cluster anions in the size range $n = 13-75$, only a few select cluster sizes followed the superatom model.

II. COMPUTATIONAL METHODS

For Al₁₂⁻¹ and Al₁₇⁻¹, the structures of Roach et al.⁹ were used as starting points for the optimized structures. The structures found by Aguado and Lopez¹⁴ were used as starting points for the minimum energy structures of Al₂₀⁻¹ and Al₂₃⁻¹, as well as confirmation of the Al₁₇⁻¹ structure. The PBE^{15,16} functional was used in the previous studies, and thus we report results with this functional to extend the previous results. In a study by Drebov and Ahlrichs,¹⁷ high-level ab initio calculations were carried out on small neutral aluminum clusters, and the results were used to test several exchange-correlation functionals. Because they did not include any meta-hybrid functionals in this study, we have extended their results by testing the M06 and M06-2X functionals of Zhao and Truhlar.¹⁸ The results are listed in Table S1 of the Supporting Information, including the cohesive energies and the dissociation energies as well as the maximum error and average error for each. The M06 functional, which has 27% exact exchange and has been parametrized to be accurate for both metallic and organic systems, has lower errors in each category than does the M06-2X functional, which has 54% exact exchange and has been parametrized primarily for main group chemistry. The M06 functional has the lowest maximum error and lowest average error of any functional for the cohesive energy, and it has the second lowest maximum error and the lowest average error for the dissociation energy; thus it was chosen for this study. Reaction paths are reported for the Al₁₂⁻¹ and Al₁₇⁻¹ systems using the PBE^{15,16} and M06¹⁸ functionals with the 6-311++G** basis set. For the Al₂₀⁻¹ and Al₂₃⁻¹ systems, results are calculated at the PBE/6-311G** and M06/6-311++G** levels of theory. All calculations were carried out with the GAMESS¹⁹ program except for the M06 calculations on Al₁₇⁻¹, Al₂₀⁻¹, and Al₂₃⁻¹, which were carried out with Gaussian 09.²⁰

The potential energy surface for each Al_{*n*}⁻¹-H₂O system was sampled by carrying out DFT local optimizations with the water molecule near each aluminum atom. Only sites with significant LUMO or LUMO+1 contribution had a binding interaction. For the Grotthuss-type mechanisms, only one additional water molecule was included for each step. While mechanisms with additional water molecules probably exist, a previous study¹¹

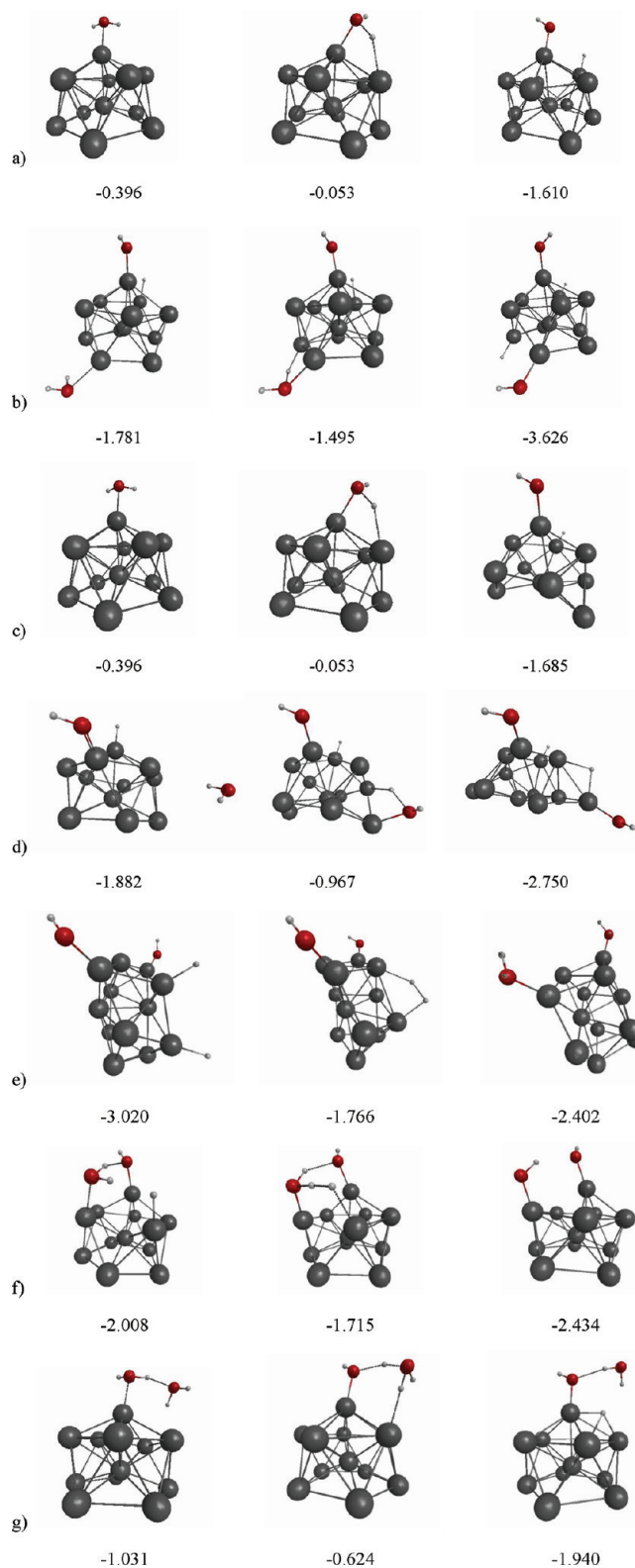


Figure 1. Reaction path structures for Al₁₂⁻¹ + water. Each row of this figure has the reactant, transition state, and product structures for the given step. The relative enthalpy, calculated with M06, is listed under each structure (eV). (a) Step 1 for LH1 and ER. (b) LH1 step 2. (c) LH2 step 1. (d) LH2 step 2. (e) LH2 step 3. (f) ER step 2. (g) GLH step 1.

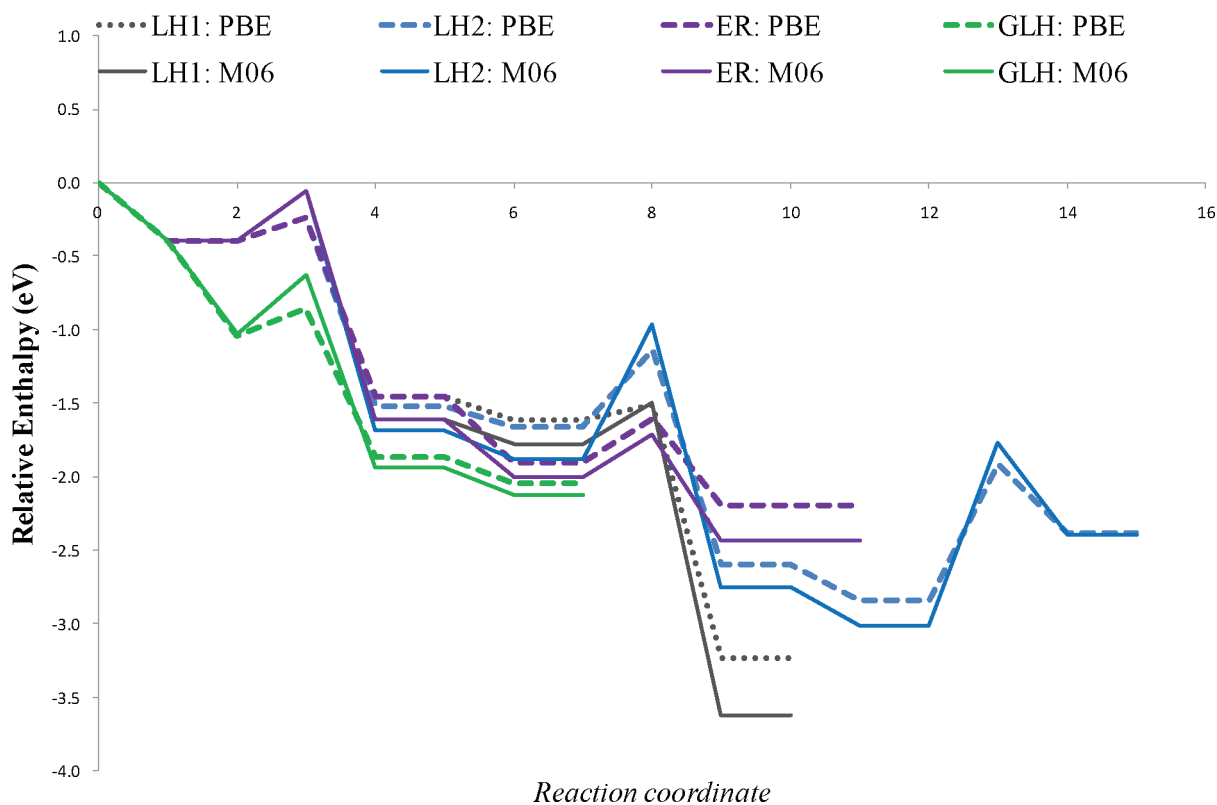


Figure 2. PBE and M06 reaction path energetics (corrected with zero-point-vibrational energies) with the 6-311++G** basis set for Al_{12}^{-1} + water. The LH1 mechanism is consistent with Roach et al.⁹ and does not release H_2 , while the GLH mechanism ends when the second water molecule does not form an Al–O bond. The other two mechanisms release H_2 .

found the lowest barrier with just one additional water molecule. The transition states for each mechanism were found by the standard search methods^{21–23} available in the GAMESS and Gaussian programs, and each transition state was confirmed to have exactly one imaginary frequency. Reaction path following calculations²⁴ were carried out to confirm that the transition state connects the reactant to the product.

III. RESULTS AND DISCUSSION

A. Al_{12}^{-1} . The extremely stable “super halide” ion Al_{13}^{-1} has an icosahedral structure with one atom in the center and the other 12 atoms located at the icosahedron’s vertices, and the Al_{12}^{-1} ion has a similar structure with one of the vertex atoms missing; thus it can be described as two stacked, staggered pentagonal rings with one atom on top and one central atom.

The LUMO for the Al_{12}^{-1} cluster has a large, protruding lobe on the “top” atom, and thus the oxygen atom of the water molecules binds to this aluminum atom, as found previously,^{9,10} and an O–H bond is broken when one hydrogen atom binds to an adjacent aluminum atom. Reactant, transition state, and product structures for each step in each reaction mechanism studied here are shown in Figure 1, as well as the enthalpy relative to the original reactants. Figure 2 shows the energetics along each reaction path. Two different structures with similar energies were found for the product of the first step. When the geometry optimization was carried out with the pure GGA functional PBE, the structure shown as the product of step1:LH1 was found, which appears to correspond to the structure found previously.^{9,10} When the optimization was carried out with a hybrid functional,

the system rearranged to the structure shown as the product of step1:LH2, where the bottom ring of aluminum atoms, instead of resembling a pentagon, now resembles a hexagon with one atom missing. When the two structures were reoptimized at each level of theory, they were found to have similar energies, with the LH2 structure lower in energy by 0.08 eV with either the PBE or the M06 functional. When the LH1 structure is used, the reaction proceeds as found previously,^{9,10} with the second water molecule binding to a site opposite the first water molecule, and the second attached hydrogen atom being too distant from the first hydrogen atom for a viable reaction path to release H_2 . However, when the LH2 structure is used, the second water molecule binds such that when the O–H bond splits, this second attached hydrogen atom is adjacent to the first one. The barrier for this second bond-breaking step is higher in LH2 than in LH1, but with the two hydrogen atoms in closer proximity in the LH2 mechanism, a path could be found where H_2 is released.

A reaction path was also found that follows the ER type mechanism in the second step. This mechanism follows the same mechanism as LH1 for the first step. As can be seen in Figure 1, in the second step of the ER mechanism, the second water molecule directly transfers a hydrogen atom to the aluminum-bound hydrogen atom from the first step, thus creating H_2 in just two steps. Because of hydrogen bonding with the other oxygen atom, the addition of the second water molecule to the cluster in the ER mechanism has an additional stabilization energy, 0.4 eV as compared to 0.2 eV for LH1 or LH2. The ER mechanism also has a low barrier for the second step, 0.3 eV, as compared to 0.5 eV for LH2 (M06: 0.9 eV). The barrier for the second step in LH1 is 0.1 eV (M06: 0.3 eV), but this mechanism does not lead to the release of H_2 .

The molecular dynamics study of Shimajo et al.¹¹ showed that, for the first OH bond-breaking step, the Grotthuss-like mechanism with one, two, or three additional water molecules has a lower barrier than the LH mechanism (zero additional water molecules), with the lowest barrier being with one additional water molecule, followed by two additional water molecules, and then three. We found that while one additional water molecule

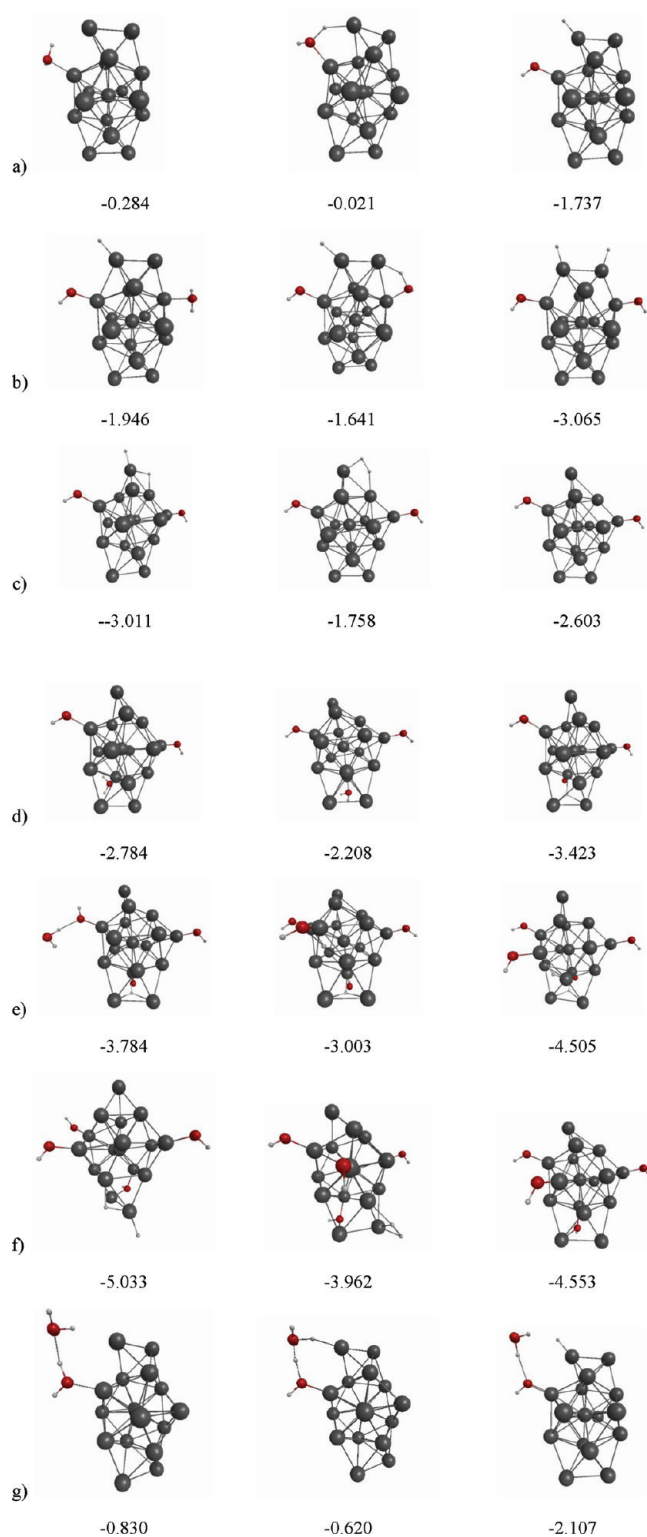


Figure 3. Continued

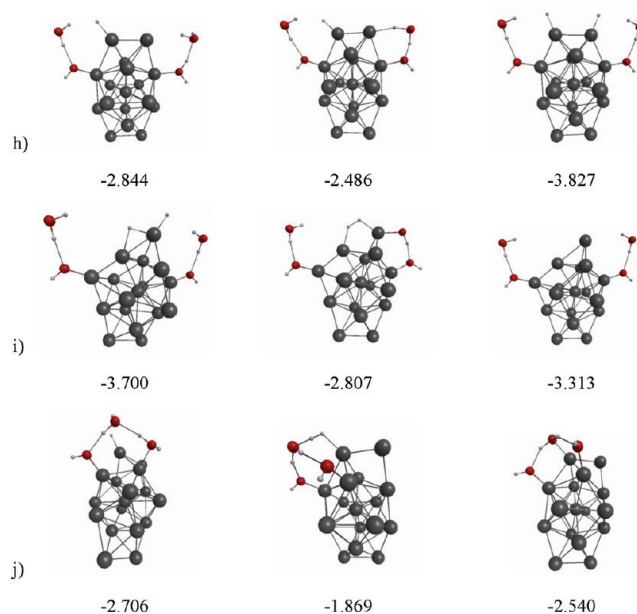


Figure 3. Reaction path structures for Al_{17}^{-1} + water. See caption for Figure 1. (a) LH step 1. (b) LH step 2. (c) LH step 3. (d) LH step 4. (e) LH step 5. (f) LH step 6. (g) GLH and GER step 1. (h) GLH step 2. (i) GLH step 3. (j) GER step 2.

provides stabilization energy of about 0.6 eV due to hydrogen bonding, the barrier for the first step is similar in both mechanisms, close to 0.2 eV (M06: 0.4 eV). However, a mechanism that completes the reaction using the Grotthuss-like mechanism could not be found for the Al_{12}^{-1} cluster. Only the first step in a possible GLH or GER mechanism was found. Thus, the ER mechanism seems most likely in this case.

Results from the PBE functional and the M06 functional are generally in good agreement. One difference is that the reaction barriers are slightly larger when the M06 functional is used. This could be expected because GGA functionals usually underestimate transition state energies. Also, for the final configuration of the LH1 mechanism, the M06 functional yields a lower energy. The particularly low energy for the final product of this mechanism, which does not release H_2 , may be the explanation for the experimental evidence that this cluster does not produce H_2 .

B. Al_{17}^{-1} . The minimum energy structure for Al_{17}^{-1} identified by Aguado and Lopez¹⁴ consists of an icosahedral Al_{13}^{-1} core with a two-atom capping bridge and an identical capping bridge opposite the first one, creating a prolate structure with nearly D_{2h} symmetry. The capping atoms are Lewis-base sites, while the adjacent “side” atoms (those coordinated to just one of the capping atoms) are the Lewis-acid sites.

Three mechanisms have been found for the release of H_2 : LH, GLH, and GER. Figure 3 gives the structure and relative enthalpy for the reactant, transition state, and product of each step in each mechanism, and the energetics are plotted in Figure 4. The two functionals are in good agreement, with the main difference being the larger barriers in the M06 calculations. An unassisted ER mechanism is not feasible due to the distance between the second Lewis-acid site and the first Lewis-base site. The LH mechanism was identified previously,^{9,10} but here we have extended it to the release of a second H_2 molecule. The two Grotthuss-like mechanisms each have one additional water

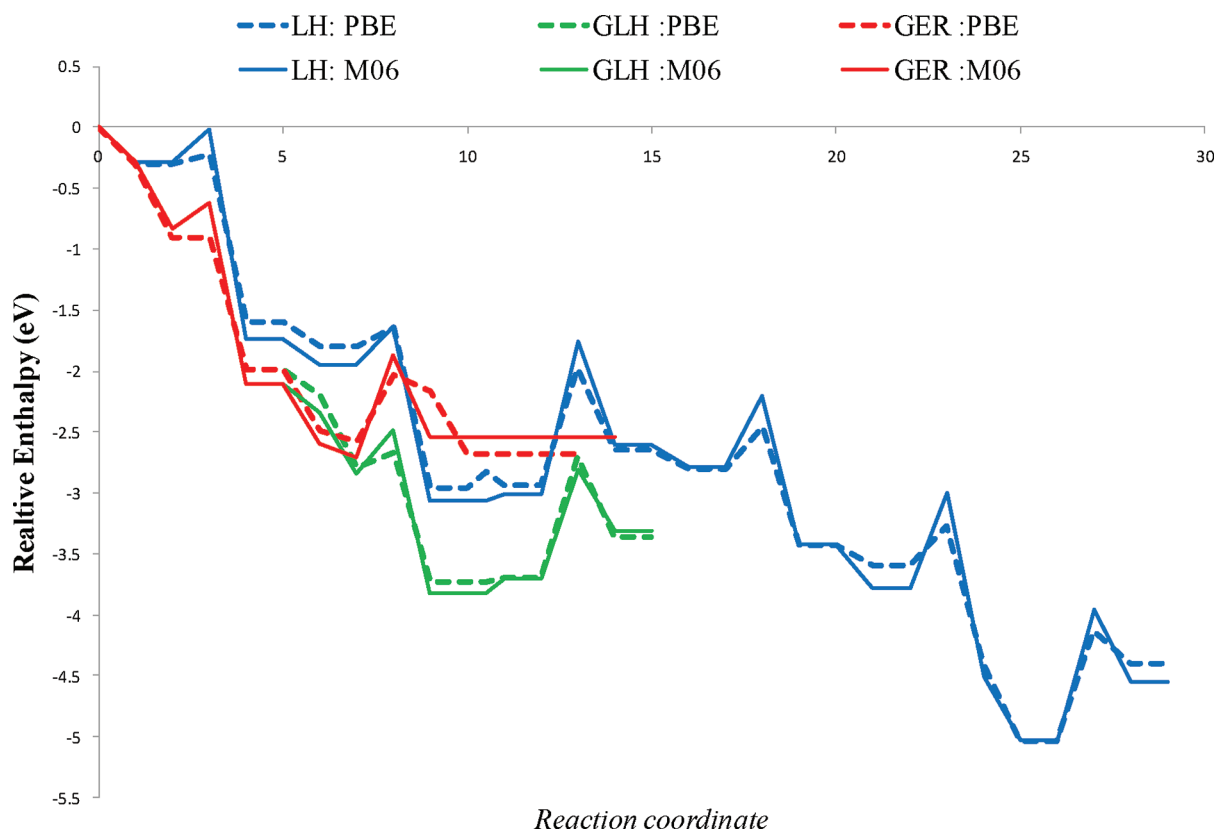


Figure 4. PBE and M06 reaction path energetics with the 6-311++G** basis set for Al_{17}^{-1} + water. The first three steps of the LH mechanism, which results in production of a H_2 molecule, are consistent with the previous results,^{9,10} but here the mechanism has been extended to the release of a second H_2 molecule. The GLH and GER mechanisms shown each release a H_2 molecule.

molecule assisting in the first O–H bond-breaking step, and appear energetically favorable, as the hydrogen bonding of the additional water provides an additional 0.6 eV in stabilization energy and the first barrier is reduced from 0.08 eV (M06: 0.26 eV) to 0.00 eV (M06: 0.21 eV). While the reaction path for the release of the H_2 molecule by the LH mechanism involves adsorption of a second water molecule with an exothermicity of 0.2 eV, in the GLH mechanism the third and fourth water molecules are adsorbed with an exothermicity of 0.8 eV, and in the GER mechanism the third water molecule is adsorbed with an exothermicity of 0.6 eV. The LH and GLH mechanisms each have a second step with a small barrier around 0.1 eV (M06: 0.3 eV) and an exothermicity around 1 eV, followed by a third step with a barrier of about 1.0 eV and an endothermicity of 0.3 eV (M06: 0.4 eV). The GER mechanism has only two steps, with the second and final step having a barrier of 0.54 eV (M06: 0.84 eV) and a small exothermicity of about 0.1 eV. While the GER mechanism might be preferred as part of a renewable energy cycle, where the lower overall change in energy results in a faster reaction rate,²⁵ the GLH mechanism's lower barrier for the second step, as well as the larger overall exothermicity, makes it appear energetically favorable. Also, the experimental results^{9,10} indicate that the GLH mechanism is preferred over the GER mechanism, as the peak at a mass of 535 is more prominent than a peak at 515. The GLH mechanism might be preferred if the additional heat of reaction can be utilized and if the aluminum hydroxide product is desired. A recent MD simulation by Ohmura et al.¹² shows this system producing three H_2 molecules.

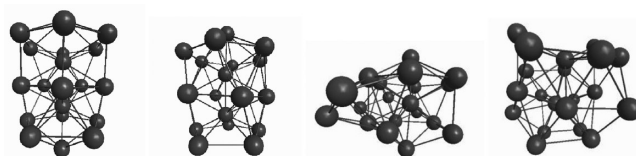


Figure 5. Structures for Al_{20}^{-1} , from left to right: M1, M2, M3, and M4.

Their mechanism for producing the first H_2 molecule appears to be an alternative GER mechanism, where the first Al–H bond is formed in a Grothuss-like step similar to that shown here, while the second step has a water molecule bind to an adjacent Al atom and transfer one of its H atoms to this bound H atom in a manner similar to the ER mechanism, but with some influence from other water molecules. Because their figures¹² do not show the full cluster or all of the water molecules involved, a full comparison with our results is not possible.

The LH mechanism has been extended for the release of a second H_2 molecule. The reaction path to release the second H_2 has higher barriers and less exothermicity than the path for the first H_2 . However, it should still be feasible because releasing the first H_2 was exothermic by 2.6 eV, and release of the second H_2 is exothermic by about 2 eV.

C. Al_{20}^{-1} . The four energy-minimized structures for $n = 20$ reported by Aguado and Lopez¹⁴ are shown in Figure 5. The first minima, labeled M1 in Figure 5, was found to be lowest in energy for the neutral, the cation, and the anion, and our calculations

Table 1. Relative Energies of Al_{20}^{-1} Structures

		M1	M2	M3	M4
M06/6-311++G**	energy (eV)	0.00	1.08	0.86	1.15
	enthalpy (eV)	0.00	1.04	0.85	1.12
PBE/6-311G**	energy (eV)	0.00	0.73	0.82	0.88
	enthalpy (eV)	0.00	0.73	0.83	0.87
PBE/DZP	ref 14	0.00	0.89	0.87	0.91

with both the PBE and the M06 functionals also found this structure to be lowest in energy for the anion, as shown in Table 1. This structure can be approximately described as the icosahedral structure of Al_{13}^{-1} with a hexagonal ring stacked on top and a single atom stacked on top of that. Thus, it has the stacking sequence 1–5–1–5–1–6–1. The structure M2 has the stacking sequence 5–1–6–1–6–1 and has an energy about 1 eV higher than that of structure M1. Structure M3 could be described as being composed of a 1–5–1–6–1 stack, with the remaining 6 atoms forming a floppy wing on one side. M3 was found to be second lowest in energy when the M06 functional was used, but still 0.8 eV less stable than M1. Structure M4 might be roughly described by the stacking sequence 4–1–7–1–7, and it has an energy 1.1 eV above the energy of M1. Four types of mechanisms have been found for the reaction of structure M1 with water to produce H_2 : LH, ER, GLH, and GER. The structures for these mechanisms are shown in Figure 6 along with the relative enthalpies, and the energetics are plotted in Figure 7. The atom in the hexagonal ring that is most evenly staggered between two atoms in the pentagonal ring below is the most electron deficient, and thus it has a protruding lobe in the LUMO. This is where the oxygen atom of a water molecule is most likely to interact. The SOMO has a large lobe on the top atom, and the OH bond can cleave across this pair of adjacent Lewis-acid–Lewis-base sites. This step is identical in the LH and ER mechanisms, while in the GLH and GER mechanisms, an additional water molecule facilitates the transfer of the hydrogen atom, significantly lowering the barrier.

In the first step for the LH and ER mechanisms, the barrier is 0.47 eV (M06: 0.80 eV), which is greater than the 0.3 eV (M06: 0.14 eV) gained by the initial coordination of the water molecule, implying a low reaction rate for these mechanisms except at high temperature. However, this step is exothermic by 1.3 eV, which should make the rest of either reaction path feasible. In the LH mechanism, coordination of a second water molecule gains another 0.44 eV (M06: 0.26 eV), while the second O–H bond-breaking barrier is 0.48 eV (M06: 0.55 eV). This step is exothermic by 0.43 eV (M06: 0.97 eV). The third step, where the H_2 leaves, has a barrier of 1.33 eV (M06: 1.77 eV), in a step that is endothermic by 0.44 eV (M06: 0.92 eV). This mechanism has an overall exothermicity of 2.44 eV (M06: 1.89 eV). In the ER mechanism, the second water attaches to an aluminum atom adjacent to the already attached OH and H, and a hydrogen bond is formed by the nonreacting hydrogen atom to the first oxygen atom, resulting in a complex slightly more stable than in the LH mechanism. The second step has a larger barrier than the second step in the LH mechanism, but is also the final step, releasing H_2 . A slight variant of the ER mechanism was also found, where the second water molecule does not form the additional hydrogen bond, making this intermediate state slightly less stable, resulting in a smaller exothermicity for its formation but then a slightly smaller barrier for the final step. The ER mechanisms seem more

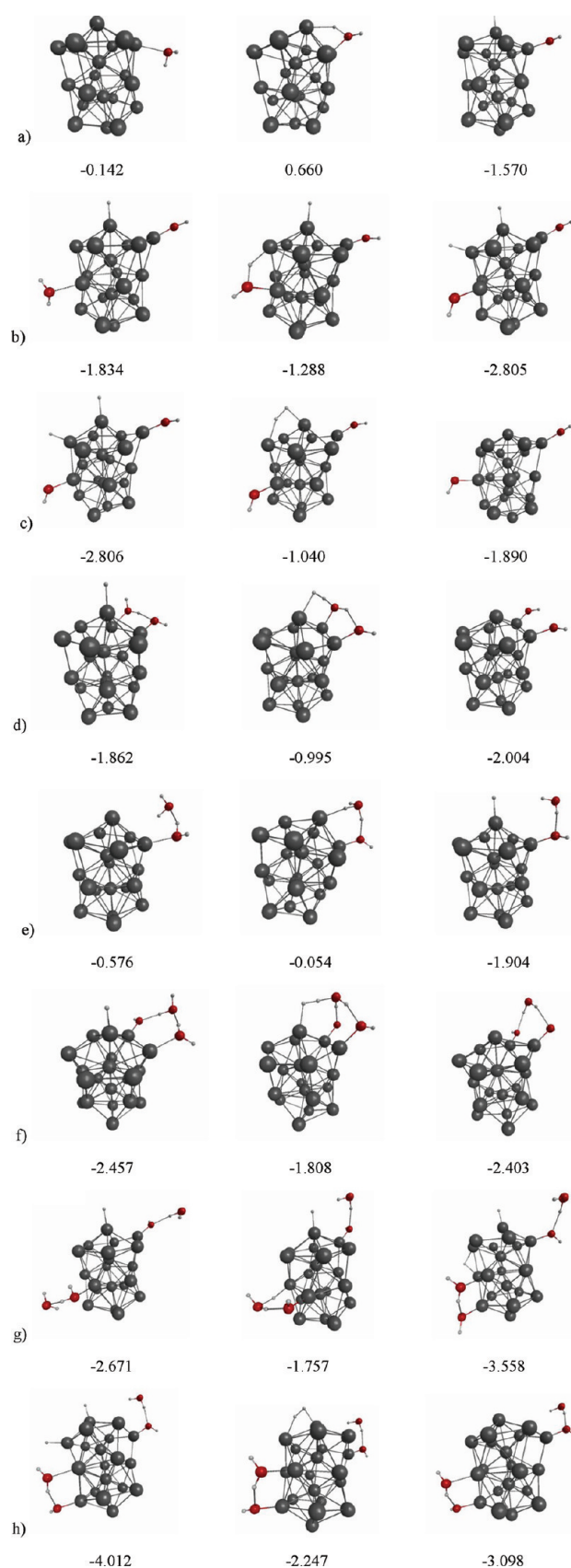


Figure 6. Reaction path structures for Al_{20}^{-1} + water. See caption for Figure 1. (a) LH and ER step 1. (b) LH step 2. (c) LH step 3. (d) ER step 2. (e) GLH and GER step 1. (f) GER step 2. (g) GLH step 2. (h) GLH step 3.

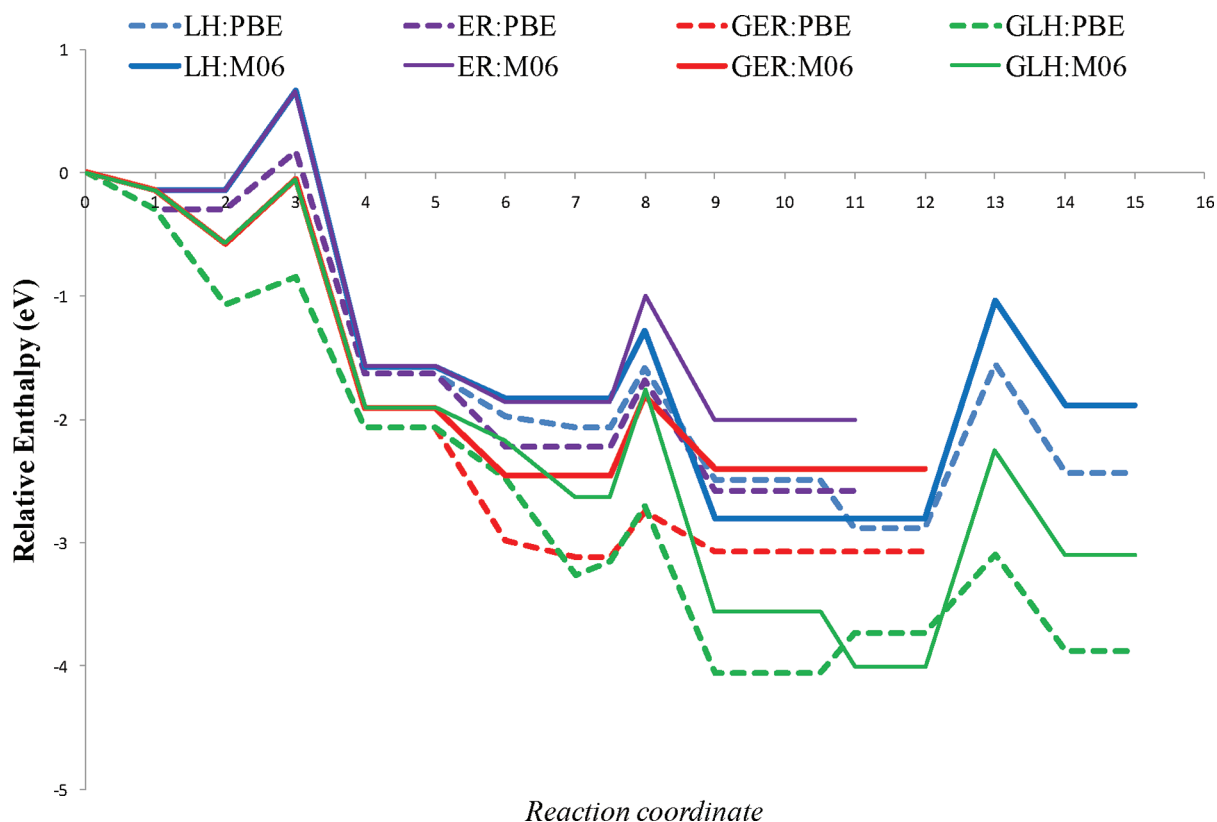


Figure 7. PBE/6-311G** and M06/6-311++G** reaction path energetics for Al_{20}^{-1} + water. The four mechanisms shown all result in the release of H_2 .

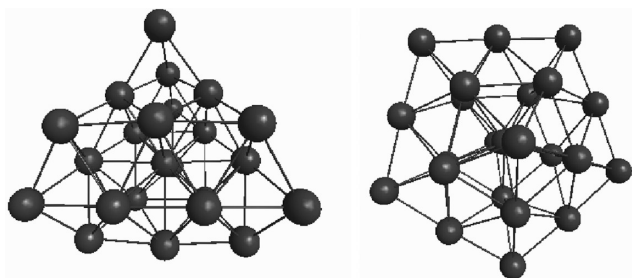


Figure 8. Structures for Al_{23}^{-1} , left to right: M2 and M3.

likely than the LH mechanism, as they avoid the third step, which has a large barrier and is endothermic.

Two reaction paths were also found for the GER mechanism, but they are mirror images of each other with no significant differences in structure or energetics. The 1.1 eV (M06: 0.6 eV) gained by the addition of the first two water molecules should be enough to get over the first bond-breaking barrier of 0.2 eV (M06: 0.5 eV). This step is exothermic by 1.0 eV (M06: 1.33 eV), and addition of the third water molecule gains another 1.1 eV (M06: 0.6 eV). The second and final step has a barrier of 0.4 eV (M06: 0.7 eV) and is approximately thermoneutral. The modest barriers and overall exothermicity of 3.1 eV (M06: 2.4 eV) make this mechanism seem quite likely.

The GLH mechanism is identical to the GER mechanism for the first step, and the barrier for its second step is similar in magnitude to the second step in the GER mechanism. However, while the second step in the GER mechanism is the H_2 -producing final step, completion of the GLH mechanism to

Table 2. Relative Energies of Al_{23}^{-1} Structures

Al_{23}^{-1}		M2	M3
M06/6-311++G**	energy (eV)	0.39	0.00
	enthalpy (eV)	0.37	0.00
PBE/6-311G**	energy (eV)	0.00	0.09
	enthalpy (eV)	0.00	0.12
PBE/DZP	ref 14	0.00	0.04

release H_2 requires a third step, which has a significant barrier. Thus, the GER mechanism seems most likely.

The experimental data for these reactions were obtained by detecting species in the product stream by mass spectroscopy.¹⁰ While the experimental report is that Al_{20}^{-1} resists reacting with water, the peak near a mass of 616 could be the GLH product,¹⁰ while the GER product at a mass of 596 may be obscured by the large peak at a mass of 594 from unreacted Al_{22}^{-1} . Because the initial binding energies for these mechanisms are small, the leaving of water before reaction occurs may be competitive with the GLH and GER mechanisms outlined here, resulting in the experimentally observed slow reaction rate, even at high water concentrations.¹⁰

D. Al_{23}^{-1} . For Al_{23}^{-1} , the second and third energy-minimized structures of Aguado and Lopez,¹⁴ labeled M2 and M3 in Figure 8, are close in energy, and while M2 is calculated to be the global minimum by 0.12 eV when the PBE functional is used, M3 is calculated to be the global minimum by 0.37 eV when the M06 functional is used, as shown in Table 2. Structure M2 might be called a highly distorted tetrahedron, while M3 is approximately a pentagonal bipyramid.

The reaction of each of these two structures with water has been investigated with the LH, ER, GLH, and GER mechanisms.

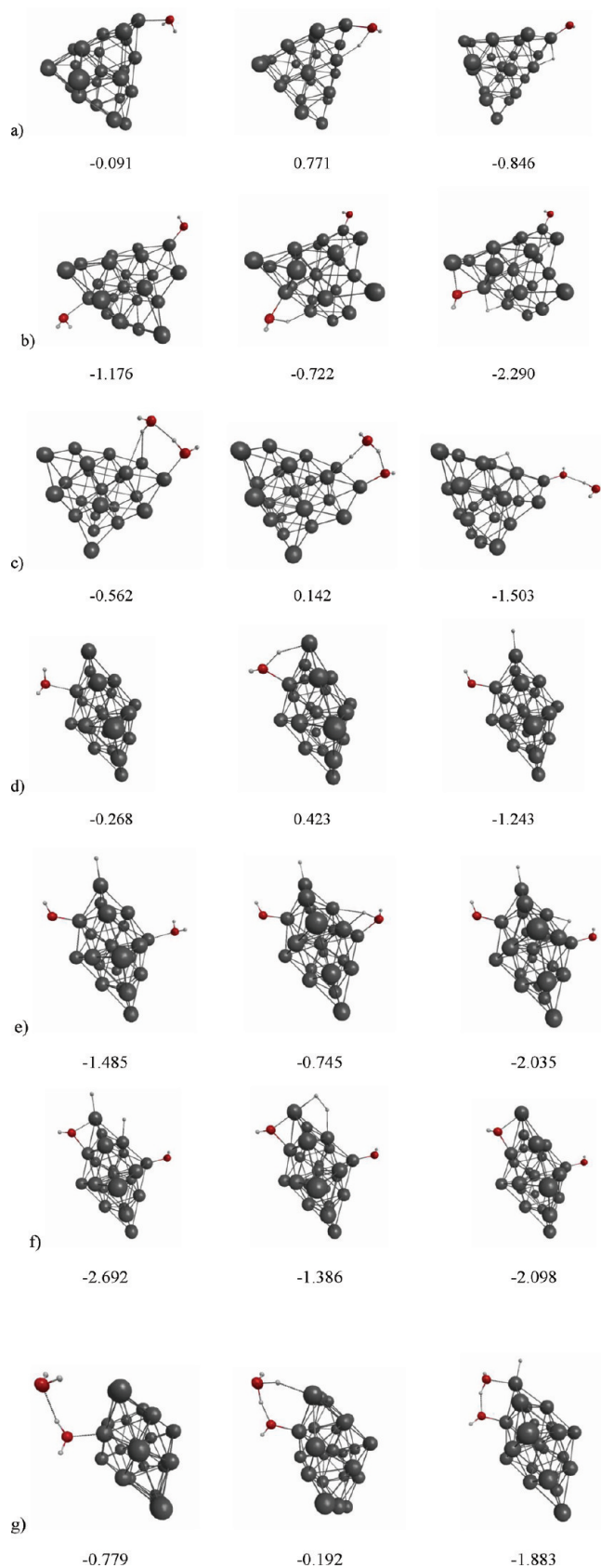


Figure 9. Continued

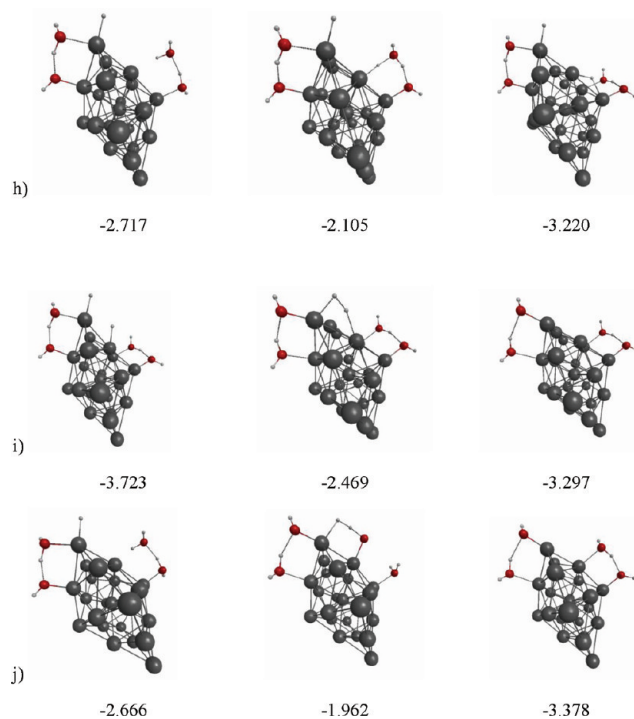


Figure 9. Reaction path structures for Al_{23}^{-1} + water. See caption for Figure 1. (a) M2 LH step 1. (b) M2 LH step 2. (c) M2 GER step 1. (d) M3 LH step 1. (e) M3 LH step 2. (f) M3 LH step 3. (g) M3 GLH and GER step 1. (h) M3 GLH step 2. (i) M3 GLH step 3. (j) M3 GER step 2.

Figure 9 shows the structures and relative enthalpies for each step in each mechanism, and the reaction energetics are plotted in Figure 10. The first two steps of the LH mechanism for structure M2 were found, with the chemisorption of the two water molecules again driven by the presence of adjacent Lewis-acid and Lewis-base sites. The resulting two hydrogen atoms bonded to the cluster in this mechanism are too far apart to form H_2 , so this mechanism fails to produce hydrogen gas. A path for the ER mechanism was not found for the M2 structure. A mechanism for the first step in a GLH or GER path was found, but a second step could not be found due to a lack of adjacent Lewis acid–Lewis base sites combined with the strong binding of the first H atom to the cluster.

When the M3 structure is used, reaction paths using the LH, GLH, and GER mechanisms were found. In the LH mechanism, the energy gained from the initial water adsorption (PBE, 0.51 eV; M06, 0.27 eV) may not be enough to surmount the barrier (PBE, 0.39 eV; M06, 0.69 eV) for the first bond splitting (particularly according to the results from the M06 functional), so this mechanism is likely to be slow except at high temperature. However, because this step is exothermic by about 1 eV, and adsorption of the second water molecule gains an additional 0.55 eV (M06: 0.24 eV), the system should have enough energy to get over the second barrier of 0.48 eV (M06: 0.74 eV). Because the second step is exothermic by 0.66 eV (M06: 0.55 eV) and is followed by a rearrangement that is exothermic by 0.35 eV (M06: 0.66 eV), the system may have enough energy to surmount the final barrier of 0.96 eV (M06: 1.31 eV). However, the two Grotthuss-type mechanisms seem more likely with their lower barriers and larger exothermicity. The GER mechanism seems particularly likely because it releases

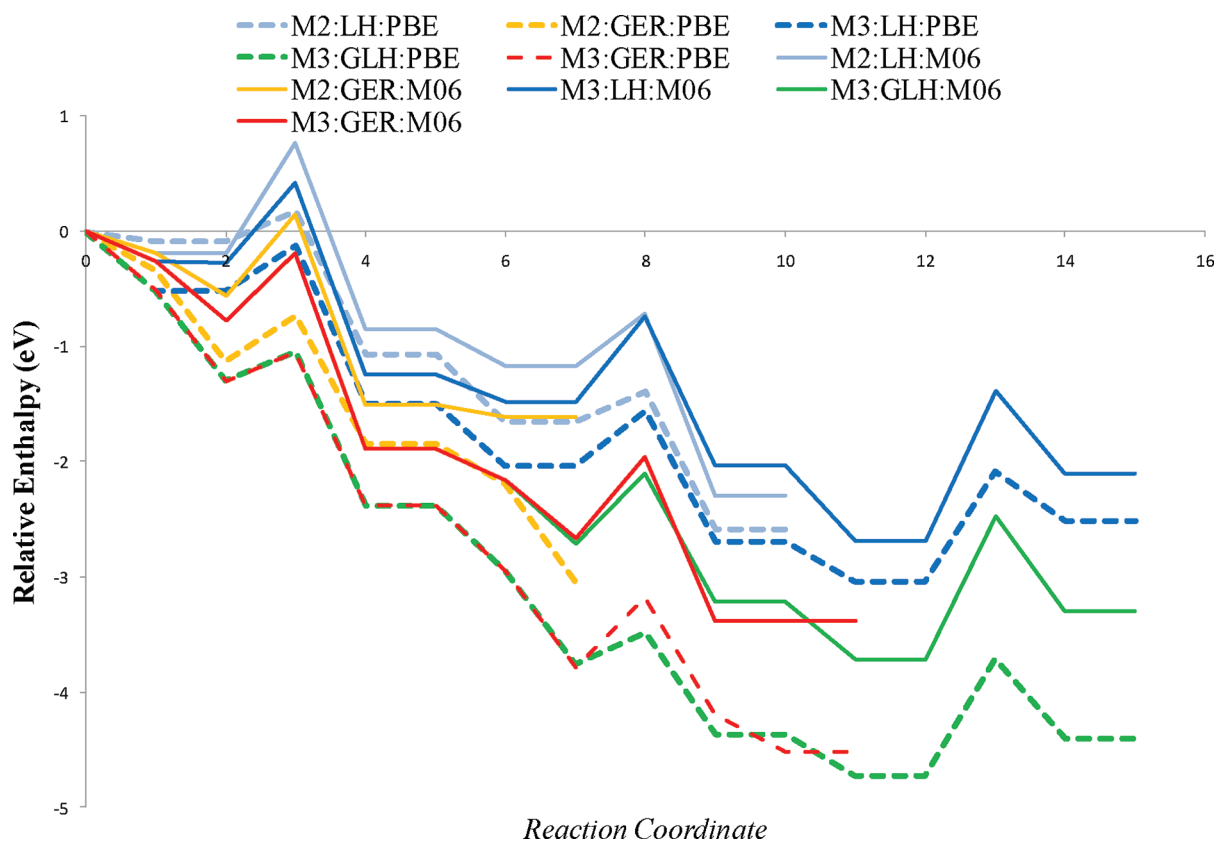


Figure 10. PBE/6-311G** and M06/6-311++G** reaction path energetics (corrected with zero-point-vibrational energies) for Al_{23}^{-1} + water. The two mechanisms starting with the M2 structure do not produce H_2 , while the three mechanisms starting from the M3 structure all produce H_2 .

H_2 after just two steps, while the third step in the GLH mechanism has a significant barrier (PBE, 1.02 eV; M06, 1.25 eV).

IV. CONCLUSIONS

The reaction of aluminum nanocluster anions with water has been studied theoretically using density functional theory. The results from the PBE and M06 functionals are in at least qualitative agreement, with the M06 calculations yielding somewhat larger reaction barriers. Our study did find significant size and structure specificity in reactivity. As found previously,¹³ the superatom model is not fully reliable for predicting the properties of aluminum cluster anions. The importance of sites that act as Lewis acids and Lewis bases, as described previously,^{9,10} was confirmed. The Grotthuss-like mechanisms, where additional water molecules facilitate the proton transfer reaction, were shown to be important paths for these reactions. For the Al_{12}^{-1} cluster, a Grotthuss-like mechanism for H_2 production could not be found. However, an ER mechanism was found, as well as a second LH mechanism that could release H_2 . The LH mechanism found previously^{9,10} was confirmed with both functionals, and due to the low barrier in its second step and particularly stable product, may be the reason for the experimental result, indicating that this cluster does not lead to H_2 production. For the Al_{17}^{-1} cluster, the previously found LH mechanism was extended to release a second H_2 molecule, and two other H_2 -producing mechanisms were found that use Grotthuss-like mechanisms. The products of these mechanisms can also be observed in the experimental data.^{9,10} For the Al_{20}^{-1} cluster, four mechanisms were found, but only the two Grotthuss-like mechanisms are likely to be

energetically feasible except at high temperature, and they may only play a major role at larger water concentrations. These factors may explain the resistance to reactivity of this cluster observed experimentally. For the Al_{23}^{-1} cluster, the initial binding energy is larger, which may explain its enhanced reactivity. The experiment did not report whether H_2 was produced in the reaction of Al_{23}^{-1} with water, yet the results of this study indicate that the M3 structure is likely to produce H_2 , but that the M2 structure might not produce H_2 even though it reacts with water. Our analysis has provided a more complete understanding of hydrogen production by Al anion clusters than previously reported, consistent with experimental results that are available so far.

■ ASSOCIATED CONTENT

S Supporting Information. Small-cluster cohesive energies and dissociation energies of Table S1, along with Cartesian coordinates for the clusters, transition states, and products in the reactions given. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: paul.day@wpafb.af.mil (P.N.D.), ruth.pachter@wpafb.af.mil (R.P.).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

We gratefully acknowledge support from AFOSR and from AFRL Section 219 funding, through a grant to Dr. Christopher Bunker, AFRL/RZ.

REFERENCES

- (1) Huang, Y.; Risha, G. A.; Yang, V.; Yetter, R. A. *Combust. Flame* **2009**, *156*, 5.
- (2) Shimojo, F.; Nakano, A.; Kalia, R. K.; Vashishta, P. *Appl. Phys. Lett.* **2009**, *95*, 043114.
- (3) Knight, W. D.; Clemenger, K.; deHeer, W. A.; Saunders, W. A.; Chou, M. Y.; Cohen, M. L. *Phys. Rev. Lett.* **1984**, *52*, 2141.
- (4) Ekardt, W. *Phys. Rev. Lett.* **1984**, *52*, 1925.
- (5) Ekardt, W. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1984**, *29*, 1558.
- (6) de Heer, W. A. *Rev. Mod. Phys.* **1993**, *65*, 611.
- (7) Leuchtner, R. E.; Harms, A. C.; Castleman, A. W. *J. Chem. Phys.* **1989**, *91*, 2753.
- (8) Martins, J. L.; Car, R.; Buttet, J. *Surf. Sci.* **1981**, *106*, 265.
- (9) Roach, P. J.; Woodward, W. H.; Castleman, A. W., Jr.; Reber, A. C.; Khanna, S. N. *Science* **2009**, *323*, 492.
- (10) Reber, A. C.; Khanna, S. N.; Roach, P. J.; Woodward, W. H.; Castleman, A. W., Jr. *J. Phys. Chem. A* **2010**, *114*, 6071.
- (11) Shimojo, F.; Ohmura, S.; Kalia, R. K.; Nakano, A.; Vashishta, P. *Phys. Rev. Lett.* **2010**, *104*, 126102.
- (12) Ohmura, S.; Shimojo, F.; Kalia, R. K.; Kunaseth, M.; Nakano, A.; Vashishta, P. *J. Chem. Phys.* **2011**, *134*, 244702.
- (13) Ma, L.; Issendorff, B. v.; Aguado, A. *J. Chem. Phys.* **2010**, *132*, 104303.
- (14) Aguado, A.; López, J. M. *J. Chem. Phys.* **2009**, *130*, 064704.
- (15) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865.
- (16) Ernzerhof, M.; Scuseria, G. E. *J. Chem. Phys.* **1999**, *110*, 5029.
- (17) Drebov, N.; Ahlrichs, R. *J. Chem. Phys.* **2011**, *134*, 124308.
- (18) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215.
- (19) 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.
- (20) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Keith, T.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, revision B.01; Gaussian, Inc.: Wallingford, CT, 2010.
- (21) Baker, J. *J. Comput. Chem.* **1986**, *7*, 385.
- (22) Helgaker, T. *Chem. Phys. Lett.* **1991**, *182*, 503.
- (23) Culot, P.; Dive, G.; Nguyen, V. H.; Ghuysen, J. M. *Theor. Chim. Acta* **1992**, *82*, 189.
- (24) Gonzalez, C.; Schlegel, H. B. *J. Chem. Phys.* **1989**, *90*, 2154.
- (25) Kozuch, S.; Shaik, S. *J. Am. Chem. Soc.* **2006**, *128*, 3355.

Improving Sampling by Exchanging Hamiltonians with Efficiently Configured Nonequilibrium Simulations

Robert M. Dirks,^{*,†} Huafeng Xu,[†] and David E. Shaw^{*,†,‡}

[†]D. E. Shaw Research, 120 W. 45th St., 39th Floor, New York, New York 10036, United States

[‡]Center for Computational Biology and Bioinformatics, Columbia University, New York, New York 10032, United States

 Supporting Information

ABSTRACT: Molecular simulations aim to sample all of the thermodynamically important states; when the sampling is inadequate, inaccuracy follows. A widely used technique to enhance sampling in simulations is Hamiltonian exchange. This technique introduces auxiliary Hamiltonians under which sampling is computationally efficient and attempts to exchange the molecular states among the auxiliary and the original Hamiltonians. The effectiveness of Hamiltonian exchange depends in part on the probability that the trial exchanges can be accepted, which involves good choices of auxiliary Hamiltonians and a good method of generating the trial exchanges. In this paper, we investigate nonequilibrium simulations as trial exchange generators and develop a theoretical model for the efficiency of Hamiltonian exchange and an algorithm to better configure such simulations. We show that properly configured nonequilibrium simulations can modestly increase the overall efficiency of Hamiltonian exchange.

1. INTRODUCTION

A major problem in many molecular simulations is inadequate sampling.^{1,2} Sampling methods such as molecular dynamics (MD) and Monte Carlo (MC) typically generate a series of molecular states by small moves, such that in the long time limit, the states are sampled with probabilities corresponding to the equilibrium distribution of the underlying Hamiltonian. In many molecular systems, however, the thermodynamically important states are separated by high energy barriers, and crossing these barriers in a simulation is rare, which, for computations of feasible length, leads to incorrect probability densities of the sampled molecular states. Inadequate sampling can be manifested in poorly converged estimates of thermodynamic averages of physical quantities; it can prevent simulations from reproducing molecular events—such as protein folding and conformational changes—observed in experiments. Poor sampling can cause inaccurate results to appear deceptively precise,³ leading to a false assumption of convergence. Many techniques have been developed to tackle the problem of inadequate sampling,¹ including the popular and powerful Hamiltonian exchange method. In this work, we propose a protocol to significantly improve the efficiency of Hamiltonian exchange and provide a systematic framework for measuring the relative benefits of the variations of Hamiltonian exchange methods.

A common approach used to enhance sampling is to introduce auxiliary Hamiltonians under which the energy barriers are significantly reduced and the sampling is efficient. The equilibrium distribution of the original Hamiltonian can be obtained either by reweighting the molecular states sampled under the auxiliary Hamiltonians⁴ or by coupling a simulation of the original Hamiltonian with simulations of the auxiliary Hamiltonians, so that the molecular states sampled under the auxiliary Hamiltonians can be “exchanged” into the simulation of the original Hamiltonian. This latter approach—with many variations—is referred to as the generalized ensemble method or

Hamiltonian exchange. In such simulations, normal MD or MC moves are interrupted by attempts to exchange molecular states generated under one Hamiltonian into the simulation under a different Hamiltonian, and such trial exchanges are accepted or rejected such that the equilibrium distribution at each Hamiltonian is preserved. Early attempts of Hamiltonian exchange, such as simulated tempering^{5,6} and parallel tempering,^{7,8} reduce enthalpic barriers by raising the simulation temperatures, in effect linearly scaling the energy function. Other Hamiltonian modifications,^{6,9,10} such as softening nonbonded interactions,^{11,12} altering dihedral terms,¹³ or reducing the effective degrees of freedom,^{14,15} have since been proposed.

In order for Hamiltonian exchange to work well, trial exchanges must be accepted with a high enough probability to allow simulations at the original Hamiltonian to benefit from the enhanced sampling at the auxiliary Hamiltonians. The efficiency of Hamiltonian exchange can be improved through the judicious selection of auxiliary Hamiltonians and through informed construction of trial exchanges. The simple trial exchange—given a molecular state under Hamiltonian H_a , switch on a different Hamiltonian H_b —has been a *de facto* choice in most reported Hamiltonian exchange simulations, in which case the only issue is to optimize the selection of auxiliary Hamiltonians. More sophisticated trial exchanges,^{16–18} in which the molecular state and the Hamiltonian are updated together in such a way that the new molecular state has increased probability in the equilibrium distribution of the new Hamiltonian, can enhance the acceptance probabilities of the trial exchanges, as illustrated in Figure 1a. Recently, nonequilibrium simulations have been proposed as a method for generating trial exchanges.^{19,20} In Hamiltonian exchange with nonequilibrium trials (HENT), both the nonequilibrium simulations in the trial exchanges and the selection of

Received: July 5, 2011

Published: December 13, 2011

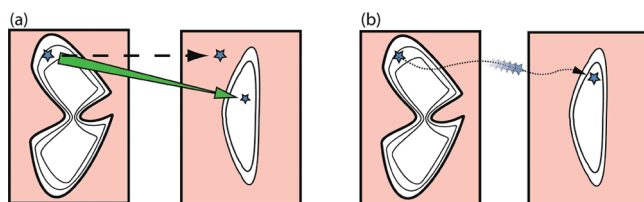


Figure 1. Benefits of sophisticated trial moves. The large rectangles represent two energy surfaces on the same two-dimensional domain, with the white and pink regions showing low and high energy regions, respectively. (a) A simple trial exchange (dotted line) keeps the coordinates (blue star) the same, producing a move unlikely to be accepted. A more efficient move (green arrow) connects the low energy regions, and has a much higher acceptance rate without the need for intermediate Hamiltonians. The changing width of the green arrow indicates a scaling of coordinates, and is accounted for by the Jacobian term in the acceptance criterion. (b) In many-atom systems, it is often unclear how to construct the most efficient trial exchanges. Instead, shortened molecular dynamics with a time-dependent Hamiltonian can be used to make a sophisticated trial exchange. This comes with an associated cost, so the challenge is to determine when the benefits outweigh the added expense. Both a and b apply to serial exchange; in replica exchange, there is a simultaneous move from the right system to the left, i.e., an arrow in the reverse direction connecting a different pair of points.

auxiliary Hamiltonians need to be planned carefully to achieve efficient exchanges. How to plan them effectively is an open question that we address in this work; our protocol optimizes the two aspects together.

To generate a trial exchange using nonequilibrium simulations, the molecular system undergoes a simulation governed by a time-dependent Hamiltonian that changes in a nonequilibrium fashion from the current Hamiltonian into the target Hamiltonian over a prescribed period of time (Figure 1b). These trial exchanges are more likely to be accepted than simple trial exchanges because the molecular state has changed gradually following the evolution of the Hamiltonian, so that at the end of the simulation, the molecular state is one that is more likely to be found in the equilibrium distribution corresponding to the new Hamiltonian. On the other hand, increased acceptance does not necessarily translate to increased computational efficiency, as the nonequilibrium simulations entail additional computational cost. The question remains whether the benefit of increased acceptance outweighs the cost of the additional simulation associated with each trial exchange. This paper lays out an objective framework to measure the overall efficiency of HENT and compares HENT to Hamiltonian exchange using simple trial exchanges.

In this work, we illustrate our method and demonstrate its effectiveness in a common application: computing the free energy associated with the transfer of a flexible molecule from the gas phase into an aqueous solution. In order to obtain the correct free energy, all of the conformations of the solute molecule must be sampled according to the Boltzmann distribution, both in the gas phase and in the solution phase, but inadequate sampling can occur when there are high energy barriers between different solute conformations. Here, we compare the effectiveness of several variations of Hamiltonian exchange in dealing with this problem. We show that at the same total computational cost, nonequilibrium trial exchanges configured using our protocol modestly enhance the overall efficiency of Hamiltonian exchange compared with methods

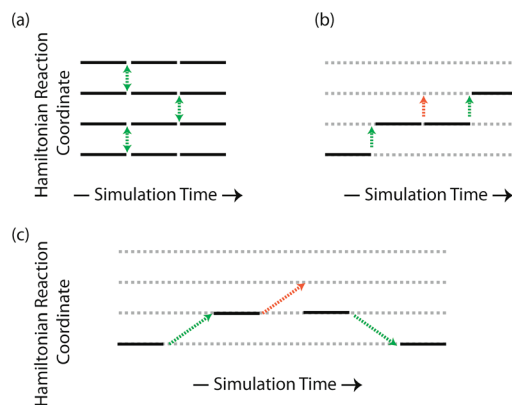


Figure 2. Three types of Hamiltonian exchange. (a) Replica exchange requires one simulation per Hamiltonian level of interest, with all simulations running simultaneously. In one common version of replica exchange, swap moves (green arrows) are attempted at regular intervals, with decisions made independently for disjoint pairs of neighbors. Swaps require no simulation time, and regardless of their outcome, there is always exactly one simulation per Hamiltonian level. (b) Serial exchange uses a single simulation that can hop between different Hamiltonian levels. If a Hamiltonian swap is rejected (red arrow), the simulation stays at the same level. One parameter per level controls the average amount of time spent simulating at each Hamiltonian. (c) Serial exchange with nonequilibrium trials replaces the direct Hamiltonian swap attempts with a simulation under a time-dependent Hamiltonian, which improves acceptance rates at the expense of additional simulation time. A similar strategy can also be applied to replica exchange (not shown).

based on simple trial exchanges, as evidenced by more accurate free energy estimates. We expect that HENT under our protocol can be a superior sampling method in general applications.

2. METHODS

2.1. Review of Hamiltonian Exchange and Nonequilibrium Simulations. Hamiltonian exchange is a technique for simulating the equilibrium distributions of a system for two or more related energy functions (Hamiltonians). By coupling multiple Hamiltonians into a single simulation, coordinates explored with one energy function can be shared with all of the other energy functions. There are two basic categories of Hamiltonian exchange simulations: replica exchange and serial exchange. Given a system with N different Hamiltonians, replica exchange methods require N simulations to be run in parallel, one per Hamiltonian. At regular or randomized intervals, the N simulations attempt to exchange energy functions (or equivalently, atomic coordinates) in some predetermined manner. A common exchange scheme⁸ for an ordered set of Hamiltonians is to have pairs of adjacent neighbors attempt Hamiltonian swaps (Figure 2a). These exchange attempts are accepted or rejected probabilistically such that the equilibrium distribution at each of the N Hamiltonians is preserved. While replica exchange requires N simultaneous simulations, serial exchange requires only one. This single simulation can traverse all N Hamiltonians by regularly attempting to change Hamiltonians (Figure 2b). Compared to replica exchange, serial exchange requires $N - 1$ additional parameters for the simulation to run efficiently.⁵ Further comparisons between serial exchange and replica exchange are included in the Supporting Information.

A sufficient condition for ensuring that exchange moves preserve the equilibrium distribution is detailed balance. This property can be defined in terms of a transformation of coordinates:

$$\pi(\hat{x}) p(T) a(\hat{x} \rightarrow T[\hat{x}]) = \pi(T[\hat{x}]) |J_T(\hat{x})| p(T^{-1}) a(T[\hat{x}] \rightarrow \hat{x}) \quad (1)$$

Here, \hat{x} is a generalized variable that includes the atomic coordinates, velocities, and other state information for all of the systems being simulated, T is a one-to-one transformation, $p(T)$ is the probability of attempting T , $J_T(\hat{x})$ is the corresponding Jacobian determinant, and $a(\hat{x} \rightarrow T[\hat{x}])$ is the probability of accepting the transformation. A simple serial exchange transformation is to keep atomic positions and velocities constant and to increment or decrement the Hamiltonian level. The Jacobian determinant of this transformation is simply 1 and is often omitted from the detailed balance equation. Individual moves that satisfy detailed balance can be strung together in a sequence to form a single combined move. While this combined move is unlikely to satisfy detailed balance, Manousiouthakis and Deem²¹ showed that such a sequence will still drive the system toward the equilibrium distribution. This justifies alternating sampling steps with a series of exchange attempts, as depicted in Figure 2a.

In molecular simulations, efficient trial moves such as those in Figure 1a are often difficult to generate. Nonequilibrium molecular dynamics can be used—in this case with a reversible, time-dependent energy function that smoothly links two target Hamiltonians^{22,23}—to generate trial moves (Figure 2c), similar to how equilibrium molecular dynamics is used to generate trial moves in hybrid Monte Carlo.²⁴ The change in the total energy of the system before and after the nonequilibrium simulation determines the acceptance probability of the trial move. In this work, such trial moves are referred to as *nonequilibrium trial exchanges*, and methods that use these simulation-based moves fall into a category referred to as *Hamiltonian exchange with nonequilibrium trials*. In contrast, direct coordinate swaps without simulation-based moves are referred to as *simple trial exchanges*, and methods that use only these moves are collectively referred to as *simple Hamiltonian exchange*.

2.2. Efficient Configuration of Hamiltonian Exchange by Minimizing Mean Round Trip Time. In a typical Hamiltonian exchange simulation, there are two energy functions of interest: a standard molecular dynamics force field meant to mimic physical properties and a modified force field meant to enhance sampling. To efficiently perform Hamiltonian exchange, one needs to have a reasonable chance of accepting exchanges between the two. In simple Hamiltonian exchange, this is traditionally accomplished by choosing a reaction coordinate connecting these two Hamiltonians and placing enough *intermediate Hamiltonians* along this path to ensure that nearest-neighbor exchanges are accepted at a reasonable rate. Much has already been written on the optimal placement of intermediates for various simple Hamiltonian exchange methods.^{7,25–28} When performing HENT, efficiently setting up the simulation becomes more complicated. Is it better to have only two Hamiltonians connected by a long nonequilibrium trial exchange, or a few intermediate Hamiltonians connected by shorter trial exchange simulations, or many intermediates without any nonequilibrium trial exchanges? A framework for answering this question is the main methodological contribution of this work and provides a way to configure efficient HENT simulations.

Scheme 1. Pseudocode for Selecting Hamiltonian Exchange Parameters

```

inputs :
   $s_i$  for all candidate levels, 1 to  $L$ 
  Candidate exchange probabilities,  $a_{i,j,t}$ 

outputs :
  Cheap paths from 1 to  $i$ ,  $\forall i \in [2, L]$ 

subroutine bestT( $i, j, \text{sumST}, \text{sumA}$ ):
  # Find locally optimal  $t$  connecting  $i, j$ 
  # Start with  $t = 0$ 
   $\text{bestScore} = 2(\text{sumST} + s_j)(\text{sumA} + 1/a_{i,j,t})$ 
   $\text{bestTij} = 0$ 
   $\text{bestAijt} = a_{i,j,0}$ 
  forall ( $i, j, t$ ) # Loop over available  $t$  for fixed ( $i, j$ )
     $\text{score} = 2(\text{sumST} + s_j + t)(\text{sumA} + 1/a_{i,j,t})$ 
    if  $\text{score} < \text{bestScore}$ 
       $\text{bestScore} = \text{score}$ 
       $\text{bestTij} = t$ 
       $\text{bestAijt} = a_{i,j,t}$ 
  return ( $\text{bestScore}, \text{sumST} + s_j + \text{bestTij}, \text{sumA} + 1/\text{bestAijt}$ )

#Main Algorithm
for  $i = 2$  to  $L$ 
  # Solve for a cheap path from 1 to  $i$ 
  # First try simple trial exchange
  ( $\text{bestScore}, \text{sumST}, \text{sumA}$ ) = bestT(1,  $i, s_i, 0$ )
  # Store  $\sum s + \sum t$  and  $\sum 1/a$ 
   $\text{bestSumST}[i] = \text{sumST}$ 
   $\text{bestSumA}[i] = \text{sumA}$ 
  for  $j = 2$  to  $i - 1$ 
    # Try adding node  $i$  to existing best path to  $j$ 
    ( $\text{score}, \text{sumST}, \text{sumA}$ ) = bestT( $j, i, \text{bestSumST}[j], \text{bestSumA}[j]$ )
    if  $\text{score} < \text{bestScore}$ 
      # New best path from 1 to  $i$ 
       $\text{bestScore} = \text{score}$ 
       $\text{bestSumST}[i] = \text{sumST}$ 
       $\text{bestSumA}[i] = \text{sumA}$ 

```

Given a predetermined reaction coordinate, one needs to decide the number and placement of intermediate levels, as well as the lengths of any nonequilibrium trial exchanges. This is a nontrivial task, for which it is necessary to have a simple quality metric to assist parameter selection. One such metric for parameter quality is the *mean round trip time* (mrtt), a measure of the average simulation time required for the system (or a particular replica of the system in the case of replica exchange) to diffuse from one end of the reaction coordinate to the other end and back again.²⁶ Conceptually, minimizing the mrtt allows the simulation to quickly diffuse to a Hamiltonian where sampling is fast, and then back to the physical Hamiltonian of interest. The main theoretical result of this paper is an inexpensive estimate of the mrtt for an N -level serial exchange process:

$$\widehat{\text{mrtt}} = 2 \left(\sum_{i=1}^N s_i + \sum_{i=1}^{N-1} t_{i,i+1} \right) \left(\sum_{i=1}^{N-1} 1/a_{i,i+1} \right) \quad (2)$$

where s_i is the amount of time spent simulating at Hamiltonian i before attempting an exchange, $a_{i,i+1}$ is the mean acceptance rate between two adjacent levels, $t_{i,i+1}$ is the length of a nonequilibrium trial exchange, and $\widehat{\text{mrtt}}$ is an estimator of the true mean round trip time. Throughout this paper, Hamiltonian exchange simulations are configured by choosing parameters, i.e., the number and placement of intermediate Hamiltonians, and the length of the nonequilibrium simulations, that minimize the value of this estimator. Equation 2 is exact for the case of a simplified Markov process with an equilibrium distribution that visits all N Hamiltonian levels equally. While it may not be

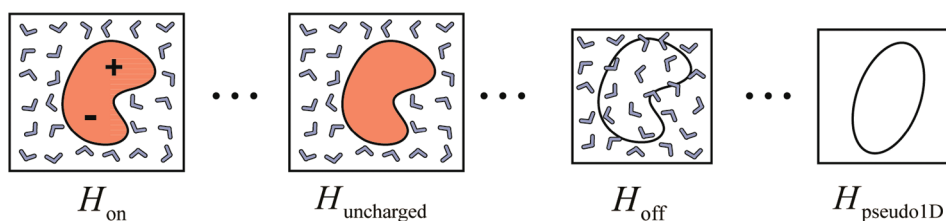


Figure 3. Hamiltonian exchange reaction coordinate. Starting with the full Hamiltonian, H_{on} , the first stage linearly scales down all ligand charges. $H_{\text{uncharged}}$ has all ligand charges set to zero but still interacts with the water through Leonard-Jones forces. The ligand–solvent interactions are then scaled down using a softcore potential, with the forces reduced to zero (white ligand) in H_{off} . The goal of the free energy calculation described here is to compute the free energy difference between H_{on} and H_{off} . To overcome energetic barriers between slowly converting conformers, the reaction coordinate is extended to H_{pseudo1D} , where the intramolecular Leonard-Jones terms in the ligand have also been removed. This greatly simplifies the ligand potential energy (depicted by the simpler shape) such that the ligand's conformations can interconvert directly, without the need for molecular dynamics. Water is not shown in H_{pseudo1D} because it has already been decoupled from the ligand. Triple dots indicate an unspecified number of intermediate Hamiltonians.

optimal to equally visit all levels,^{26,29} this is a reasonable way to simplify parameter selection. The derivation of eq 2 is provided in the Supporting Information.

In this work, the s_i values in eq 2 are fixed at 1 ps for almost all i (see section 2.3 for the exception). There is no clear consensus at this point on the optimal frequency of exchange attempts,^{30,31} the s_i values used in this study were not chosen according to any particular optimality criterion but fall well within the (rather wide) range of values that have been employed in various studies reported in the literature.^{32–34}

Equation 2 applies specifically to serial exchange but can also be adapted to replica exchange with a few small modifications. Viewing replica exchange as N loosely coupled serial exchange runs,³⁵ eq 2 still holds with the exception that all $s_i \leftarrow \max_i(s_i)$ and all $t_{i,i+1} \leftarrow \max_i(t_{i,i+1})$, due to the synchronous nature of replica exchange. The mrrt is based solely on simulation time used, and it is assumed that all other computations, such as energy evaluations, have negligible costs. This assumption breaks down for extremely small values of s_i but is reasonable for the examples presented here.

Practically, eq 2 provides a simple metric for parameter optimization, under the assumption that mrrt is a reasonable predictor of Hamiltonian exchange efficiency. Given L possible Hamiltonian levels, along with estimated $a_{i,j}$ for given $t_{i,j}$, a simple dynamic program³⁶ can quickly choose an mrrt-optimal subset of these intermediates, along with optimal lengths for nonequilibrium trial exchanges. (A method for estimating $a_{i,j}$ as a function of $t_{i,j}$ is described in section 2.5.) The total number of levels selected is an output of the algorithm and need not be specified beforehand. This optimization algorithm is similar to finding the minimum cost path connecting points 1 and L , given $L - 2$ possible intermediates and various pairwise edge costs. The problem here is slightly more difficult, as the total cost of a path is not simply the sum of independent edge costs, but similar ideas can still be used. A simple, near-optimal algorithm is given by the pseudocode in Scheme 1. While this algorithm is likely sufficient for practical purposes, refinements to guarantee optimality are discussed in the Supporting Information.

2.3. Alchemical Free Energy Calculations with Hamiltonian Exchange. Hamiltonian exchange can be used to compute the free energy required to transfer a flexible small molecule from the gas phase to an aqueous environment.³⁷ One part of this calculation³⁸ is the conversion of a fully solvated molecule to an uncharged small molecule with no interaction with the solvent. These two states are depicted in Figure 3 as Hamiltonians H_{on}

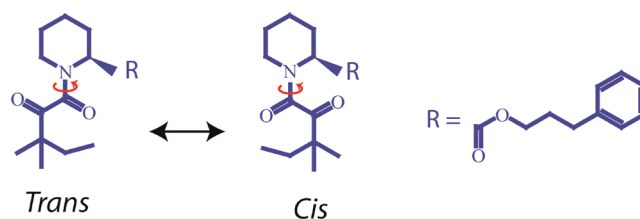


Figure 4. The FKBP ligand used in this study. Under standard conditions, this ligand exists as a mixture of *trans* and *cis* isomers. A large potential energy barrier prevents rapid rotation about the amide bond.

and H_{off} , respectively. Although Hamiltonian H_{off} should provide some enhanced sampling by removing the effects of solvent viscosity on the small molecule, this effect is small. To speed up sampling further, an auxiliary Hamiltonian, H_{pseudo1D} , is introduced and allows for faster conformational changes (see below).

The small molecule studied here is a ligand for FK506-binding protein (Figure 4). This ligand is one of several that have previously been examined³⁹ and is known to have at least two slowly interconverting conformers. The experimental conversion time between these forms is on the order of milliseconds to seconds, and sampling of both states by plain molecular dynamics is unlikely on the time scales typically used for free energy calculations.

There are several methods for overcoming energy barriers^{40–42} that could be used to facilitate transitions among the slowly converting conformers. The method that we use in this work extends the reaction coordinate to a new Hamiltonian, H_{pseudo1D} , where the ligand is not only decoupled from the solvent but all ligand–ligand nonbonded interactions have also been turned off. As a result, the potential energy of a ligand under H_{pseudo1D} is simply the sum of bond, angle, and dihedral terms. If bond lengths and bond angles are temporarily fixed, each rotatable bond in the molecule, with the exception of those in a ring, has an energy function that is independent of all the others, and the one-dimensional probability distributions around each bond can be easily computed. This decomposition into independent, one-dimensional problems allows for direct resampling of all non-cyclic rotatable bonds using standard numerical techniques such as acceptance–rejection sampling.⁴³ While bond lengths and angles are not varied with this type of move, molecular dynamics at other Hamiltonian levels should provide sufficient sampling for these other degrees of freedom. Thus, when a serial exchange

simulation reaches H_{pseudo1D} , plain molecular dynamics is replaced with a direct resampling of rotatable bonds, and $s_{\text{pseudo1D}} = 0$. In replica exchange, the same direct resampling can be performed, but since all other levels are still performing molecular dynamics, there are no wall-clock savings by skipping the simulation step only at this level. Thus, when performing replica exchange at H_{pseudo1D} , plain molecular dynamics is used along with direct resampling of all noncyclic rotatable bonds.

2.4. Reaction Coordinate Details. To connect H_{on} and H_{off} , a reaction coordinate⁴⁴ is chosen (Figure 3) that first linearly scales the charges to zero, yielding an intermediate Hamiltonian, $H_{\text{uncharged}}$. Next, the ligand–solvent Leonard-Jones interactions are turned off with a softcore potential⁴² (see Supporting Information), described by a parameter λ_{vdw} . When $\lambda_{\text{vdw}} = 0$, this is the standard Leonard-Jones potential, and when $\lambda_{\text{vdw}} = 1$, this potential is zero everywhere. The segment of the reaction coordinate joining $H_{\text{uncharged}}$ to H_{off} linearly scales λ_{vdw} from zero to one. In order to overcome energy barriers, the reaction coordinate is further extended to H_{pseudo1D} by turning off all ligand–ligand Leonard-Jones interactions, using a similar softcore potential.

2.5. Initialization Data and Input Estimation. The results of the parameter selection algorithm will only be as good as the $a_{i,j,t}$ supplied to it. To estimate these $a_{i,j,t}$, the reaction coordinate is first subdivided into a large number of possible intermediates. In the reaction coordinate segment connecting H_{on} and $H_{\text{uncharged}}$, there are 21 candidate intermediates for the charge scaling parameter, ranging from one to zero. In the region $H_{\text{uncharged}}$ to H_{off} , there are 51 candidate intermediates for the ligand–solvent softcore parameter, ranging from zero (full strength) to one (completely off). In the final segment from H_{off} to H_{pseudo1D} , ligand–ligand interactions are disabled with 21 candidate intermediates. Such a fine subdivision would not be necessary for actual applications, but for method comparison purposes, a conservative approach is taken. Further detail is provided in the Supporting Information.

Using the candidate intermediates described above, a series of 300 ps molecular dynamics runs are performed, starting with H_{pseudo1D} and ending with H_{on} . In the first leg, H_{pseudo1D} to H_{off} , water is unnecessary, as it is decoupled from the ligand, and the ligand is simulated in an otherwise empty box and run with a constant volume, constant energy integrator. Every 1 ps, the velocities are resampled to generate a constant volume, constant temperature ensemble. An equilibrated water box is then overlaid onto the ligand, and the remaining two legs are run analogously to the first, but with an NPH integrator, and velocity resampling (including an Andersen piston⁴⁵) to maintain an NPT ensemble. After the first 60 ps of each run, the potential energy of the system is periodically evaluated at all other Hamiltonians within the leg. From these data, relative free energy estimates are made for all levels, using Bennett analysis⁴⁶ or multi-Bennett analysis.⁴⁷ (In this study, only minor differences exist between standard and multi-Bennett, so for the remainder of the work, standard Bennett analysis is used, as it is computationally cheaper.) These relative free energy estimates directly give the necessary parameters needed for serial exchange.^{5,48} Furthermore, these same data are used with the free energy estimates to approximate the mean acceptance rates, $a_{i,j}$, of direct exchanges ($t_{i,j} = 0$) between all pairs of levels within a leg. These same data are also analyzed to estimate replica exchange acceptance rates. In simple Hamiltonian exchange, these acceptance rates are sufficient for mrvt-optimal level placement, using the minimum-cost path algorithm outlined above.

To estimate the impact of nonequilibrium trial exchanges, additional simulations are needed for pairs of possible intermediates. Exhaustive testing of all pairs is not practical, so only a limited set of pairs is explored. These tests involve running two-state serial exchange runs for select pairs, with separate simulations needed for each $t_{i,j} \neq 0$ of interest. Analyzing the work values⁴⁹ from these simulations gives estimated mean acceptance rates as a function of simulation length. The choice of which (i,j) pairs and $t_{i,j}$ to test is done by hand, and the specific cases examined are shown in the Supporting Information. In the end, the parameter selection algorithm can only choose nonequilibrium trial exchanges that are explicitly tested. Including additional nonequilibrium exchanges might further improve the performance of these methods relative to simple serial exchange.

One final aspect of initialization leverages the use of H_{pseudo1D} . While the initialization procedure outlined above may work well for a single ligand conformation, it may miss other important conformers. To account for this, the above procedure is repeated multiple times. Starting with a snapshot of the system under Hamiltonian H_{pseudo1D} , the system is quickly minimized by running 60 ps of constant volume, constant temperature molecular dynamics at 5 K. From this minimized structure, the equilibrium probability densities for rotation around each noncyclic rotatable bond are independently analyzed and divided into rotamers, using a $5 k_B T$ minimum peak to trough distance as a rotamer definition. The free energies of these rotamers are then estimated (see Supporting Information), and the free energy of a conformer is approximated as the sum of its constituent rotamer energies. Here, the six lowest free energy conformers (with Hamiltonian H_{pseudo1D}) include three *trans* and three *cis* rotamers (Figure 4) and are chosen for further analysis. The initialization procedure from the preceding paragraphs is performed on all six of these conformers, and the data are combined in a straightforward way to give revised input parameters, appropriately averaged (see Supporting Information) across all tested conformers. The nonequilibrium trial exchange tests are only performed at the end, once the composite free energy estimates are made using all six conformers. This rather elaborate procedure is unnecessary for the correctness of the method, but helps in selecting parameters for fair comparisons among different variants of Hamiltonian exchange.

2.6. Production Data and Analysis. Once all of the necessary parameters are chosen by the optimization algorithm, production data are generated for three methods: simple replica exchange, simple serial exchange, and serial exchange with nonequilibrium trials. The optimized replica exchange simulation uses 19 Hamiltonian levels and is run for 200 ns. The starting conformations for each level are picked from the initialization simulations and include a roughly equal number of structures from each of the six selected conformers. In both versions of serial exchange, 19 independent runs of 100 ns are performed using the same starting coordinates as in replica exchange. The simulation costs of any nonequilibrium trial exchanges are included in the total run time. When comparing the three methods, only the first 100 ns of replica exchange are used. The additional 100 ns of replica exchange are used to help create a gold standard computational result (see Results). Free energies for these long simulations are computed using versions of Bennett's method,^{44,47,50} with details given in the Supporting Information.

Mean absolute errors in free energy estimates are computed for varying amounts of total simulation time, which for HENT includes the simulation time used to generate nonequilibrium

Table 1. Estimated Free Energy Differences and 90% Confidence Intervals

method	total simulation time	relative free energy (kcal/mol @ 300 K)
simple replica exchange (data from 19 replicas lumped together)	3.8 μ s	-55.41 (gold standard)
simple replica exchange (first half only, data analyzed as 19 independent trajectories)	1.9 μ s	-55.41 \pm 0.05
simple serial exchange (19 independent runs)	1.9 μ s	-55.47 \pm 0.05
serial exchange with nonequilibrium trials (19 independent runs)	1.9 μ s	-55.40 \pm 0.03

trial exchanges. Each simulation is divided into nonoverlapping time intervals, and free energies are estimated for each time interval. The differences between these estimates and the gold standard are averaged across all of the time intervals to obtain the mean absolute error. The 90% confidence intervals are calculated subject to the assumption that the data in different time intervals are independent. This assumption is justified when the correlation between neighboring intervals is small; here, the total simulation time is comparable to the mrtt, and the data are gathered across 19 replicas.

2.7. Molecular Dynamics Details. Simulations are run using the molecular dynamics package, Desmond,⁵¹ using periodic boundary conditions, with long-range electrostatics handled by particle mesh Ewald.⁵² Calculations are performed with a reference temperature of 300 K and a reference pressure of 1 bar. A custom plugin is used to handle Hamiltonian exchange. Simulations use a 1:1:3 RESPA schedule with an inner time step of 2 fs, and bonds involving hydrogen atoms are constrained using a numerically superior implementation⁵³ of M-SHAKE.⁵⁴ Additional molecular dynamics and Hamiltonian exchange details are provided in the Supporting Information.

3. RESULTS AND DISCUSSION

3.1. Method Validation. To validate the correctness of nonequilibrium simulations as trial exchange generators, the relative free energy defined in section 2.3 is computed using three different types of Hamiltonian exchange. The two control methods, simple replica exchange and simple serial exchange, hold system coordinates constant while attempting Hamiltonian exchanges. The third method, serial exchange with nonequilibrium trials, adjusts system coordinates via molecular dynamics in order to improve acceptance rates. Table 1 shows that all three methods produce statistically equivalent free energy differences. The first row of the table gives the free energy estimate from replica exchange, using all 3.8 μ s of aggregate simulation time. Since simple replica exchange is the most commonly used Hamiltonian exchange method, this free energy estimate is used as the gold standard for all future analysis. Other choices of the gold standard, such as a maximum-likelihood weighted average of all three methods' results, yield similar results (a difference of only 0.01 kcal/mol in the gold standard) to those presented in this paper.

3.2. Efficiency of Nonequilibrium Trial Exchanges. As a measure of the efficiency of nonequilibrium trial exchanges, the mean absolute errors in free energy estimates are compared for varying amounts of total simulation time. Figure 5 graphs these errors for the three Hamiltonian exchange methods described above. Serial exchange with nonequilibrium trials is the clear winner and provides a roughly 20% reduction in mean absolute error for all time lengths considered.

Less direct measures of efficiencies are the estimated and observed mean round trip times for each Hamiltonian exchange

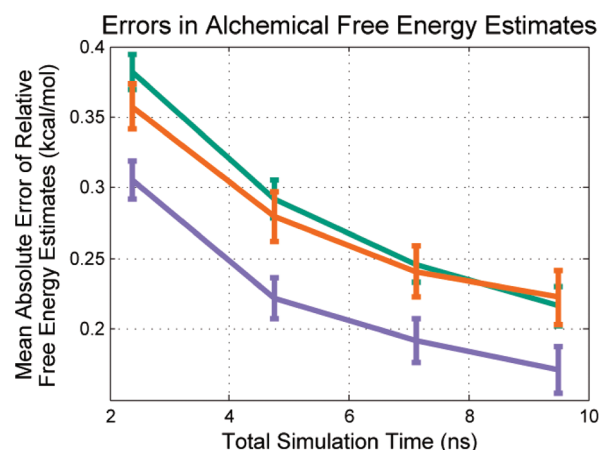


Figure 5. Error reductions from nonequilibrium trial exchanges. The mean absolute error (relative to a long replica exchange control) is shown as a function of total simulation time used. In simple replica exchange (green), this involved a single run with 19 replicas, while the two serial exchange methods (orange, purple) combined data from 19 independent runs each. Simple serial exchange (orange) performed comparably to simple replica exchange (green), while serial exchange with nonequilibrium trials (purple) reduced the average error by roughly 20% across the board. Error bars represent 90% confidence intervals.

simulation. As described previously, the estimated mean round trip time is the quality metric used to choose the number and placement of intermediate Hamiltonians along the reaction coordinate, as well as the lengths of any nonequilibrium exchange simulations. Table 2 shows the results of minimizing the estimated round trip times, using short initialization data and the algorithm outlined in section 2.2. Comparing simple replica exchange and simple serial exchange shows that the latter uses fewer intermediates and is predicted by eq 2 to reduce the mean round trip time by a factor of 2. Allowing for nonequilibrium trial exchanges further reduces both the number of intermediates and the estimated mean round trip time. Actual mean round trip times for the production simulations are shown alongside the predicted values, and the relative speeds of these three methods appear to be accurately captured by the model. This correlation between estimated and observed mean round trip times suggests that the parameter selection algorithm described here is a reasonable way to set up and compare different Hamiltonian exchange methods. Replica exchange with nonequilibrium trials was not directly tested, but the relative benefits of nonequilibrium simulations will likely be similar to those seen for serial exchange. Finally, although simple serial exchange has shorter mrtt than replica exchange, the former does not appear to significantly outperform the latter with respect to the mean absolute error of free energy estimates (Figure 5). Thus, mrtt alone is not always a perfect measure of sampling efficiency.

Table 2. Theoretical and Observed Mean Round Trip Times and Speedup Factors Relative to Replica Exchange

method	optimal number of levels	theoretical time (ps)	theoretical speedup factor	observed time with 90% confidence intervals (ps)	observed speedup factor
simple replica exchange	19	3189		4423 ± 156	
simple serial exchange	14	1529	2.1 x	1967 ± 65	2.2×
serial exchange with nonequilibrium trials	8	993	3.2 x	1286 ± 34	3.4×

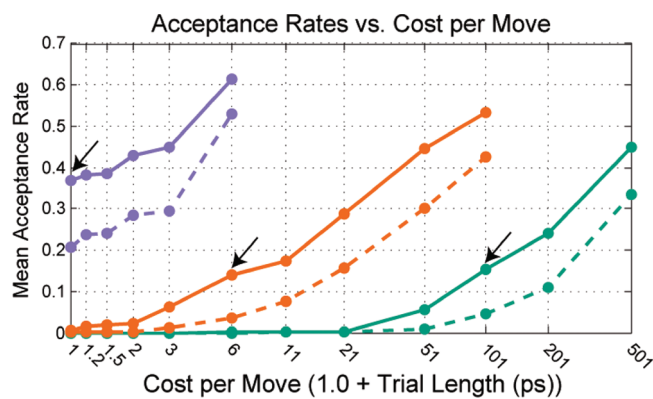


Figure 6. Approximating the efficiency of nonequilibrium trial exchanges. All curves show estimates of the mean acceptance rate for changing the Leonard-Jones interactions between an uncharged ligand and the solvent molecules, with solid lines indicating serial exchange and dashed lines indicating replica exchange. The purple curves represent an easy case: switching between a softcore coefficient of 0.88 and 1.0. The orange curves switch between 0.82 and 1.0, and the green curves switch between 0.0 and 1.0. The arrows indicate the point where the mean acceptance rate divided by the cost per move is maximized for serial exchange. In the hardest case (green), the maximum ratio occurs for 100 ps trial moves, whereas for the easiest case (purple), the maximum occurs with direct swaps between levels. These data were generated by performing two-level serial exchange simulations on a single conformer of the FKBP ligand. Replica exchange estimates were made based on these same data.

3.3. Effects of Nonequilibrium Simulation Lengths and Intermediate Hamiltonian Placement. Estimates of the mean acceptance rates for select nonequilibrium exchange attempts are shown in Figure 6. In all cases, as the nonequilibrium simulation time increases, the mean acceptance rates go up. Black arrows in Figure 6 indicate the point at which the *efficiency ratio*, defined here as the mean acceptance rate divided by the move generation cost, is at a maximum. Clearly, there is no single optimal length for nonequilibrium trial exchanges, as the answer depends on the location of the two end points along the reaction coordinate. In cases where acceptance rates are already high with a simple trial exchange, performing a nonequilibrium simulation decreases the efficiency ratio. When acceptance rates are extremely low for simple swaps, adding nonequilibrium trial exchanges can significantly boost the efficiency. Looking at the green curve, the efficiency ratio increases by a factor greater than 10^{10} when comparing 100 ps nonequilibrium trial exchanges to simple trial exchanges. This dramatic improvement is highly misleading, as someone setting up a simple Hamiltonian exchange simulation would never choose adjacent levels so far apart that the mean acceptance rates were near zero. Instead, more intermediate levels would be introduced, drastically reducing the benefit of nonequilibrium trial exchanges for any pair of adjacent levels.

Table 3. Hamiltonian Levels and Parameters for Serial Exchange with Nonequilibrium Trials

level	ligand charge coefficient	softcore λ_{vdw} (ligand-solvent)	softcore λ_{vdw} (ligand-ligand)	$t_{i,i+1}$ (ps)
0 (H_{on})	1	0.000	0.0	2
1 ($H_{\text{uncharged}}$)	0	0.000	0.0	5
2	0	0.775	0.0	0
3	0	0.825	0.0	0
4	0	0.875	0.0	0
5 (H_{off})	0	1.000	0.0	0.5
6	0	1.000	0.7	0
7 (H_{pseudo1D})	0	1.000	1.0	

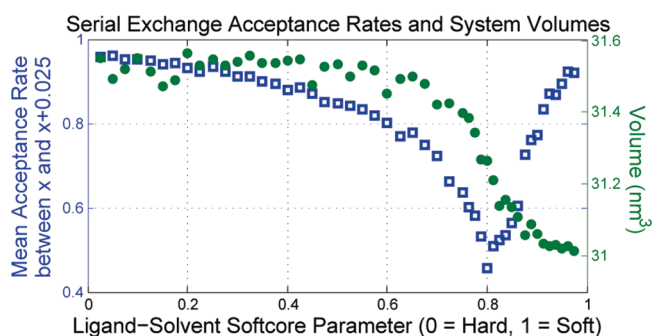


Figure 7. Trial move acceptance is not uniform. Using the initialization data from a single conformer of the FKBP ligand, the estimated mean acceptance rate (blue squares) of equidistant swaps, and the estimated mean box volume (green circles) are plotted as a function of the reaction coordinate. When x equals zero, the ligand-solvent Leonard-Jones interactions are at full strength, whereas at $x = 1$, these terms are zero. Mean acceptance rates are not uniform along the reaction coordinate, highlighting the need to carefully distribute intermediate levels. The lowest mean acceptance rates occur near the inflection point of the volume curve.

This example highlights why it would be unfair to simply choose a single set of Hamiltonian levels as the basis for comparison between simple Hamiltonian exchange and Hamiltonian exchange with nonequilibrium trials. Instead, eq 2 is used as a basis to choose reasonable parameters for each method. Along these lines, Table 3 shows the location and nonequilibrium simulation times for the eight Hamiltonian levels used with serial exchange. Of the seven possible nonequilibrium trial exchanges, four are chosen to be instantaneous (i.e., the same as simple serial exchange), and the remaining three have lengths of 0.5, 2, and 5 ps. The details of the Hamiltonian levels selected for the other two methods are given in the Supporting Information.

Figure 7 shows both the average simulation box volume and the estimated acceptance rates of serial exchanges without

Table 4. Estimated Mean Acceptance Rates for Simple Trial Exchanges between a Fully Charged and a Partially Charged Ligand

ligand charge coefficient	serial exchange acceptance rate	replica exchange acceptance rate
0.95	0.82	0.74
0.85	0.48	0.32
0.75	0.26	0.11
0.55	0.12	0.03
0.45	0.05	<0.01

nonequilibrium trials, plotted as functions of the reaction coordinate. The shape of the acceptance rate curve shows that equal size moves along the reaction coordinate do not have an equal chance of success, reinforcing the need for care when choosing intermediate level placement, even when not using nonequilibrium trial exchanges. Strikingly, the acceptance rates dip to their lowest point just when the volume of the box changes the most rapidly. This volume reduction corresponds to a phase change, wherein the ligand–solvent interaction has become soft enough that solvent molecules can stably occupy the same physical space as the ligand. The corresponding drop in acceptance rates supports the assertion that additional intermediate levels are needed close to phase transitions.^{7,26}

Table 4 compares estimated mean acceptance rates of simple replica exchange and simple serial exchange for different amounts of charge scaling. In all cases, not just those shown in Table 4, the estimated mean acceptance rates were higher for serial exchange, a result consistent with work on temperature-based exchange methods.⁴⁶

4. CONCLUSIONS

In this work, a method is developed for efficiently configuring HENT. In the computation of solvation free energy studied here, the introduction of nonequilibrium trial exchanges reduces the mean absolute error of free energy estimates, without increasing the overall simulation cost. This method can be readily adapted to problems besides free energy calculations and is applicable wherever Hamiltonian exchange is employed.

The observed reductions in statistical error reported here are encouraging, yet several additional factors should be considered. First, the comparisons described in Figure 5 exclude the cost of parameter selection, as the initialization data generated are likely overkill for typical applications. While the use of nonequilibrium simulations does require extra initialization runs to estimate the benefits of nonequilibrium trial exchanges, it is unclear how much this would actually cost. If general rules could be developed for the most likely placement and length of nonequilibrium simulations, the cost of testing them may be minimal. Furthermore, if production runs are significantly longer than the fixed cost of initialization—and they usually are—this cost will be negligible. A caveat is that our method chooses parameters to minimize estimated mean round trip times, which does not guarantee minimized sampling error. One important parameter is the simulation interval between Hamiltonian exchange attempts, fixed here as $s = 1$ ps. Equation 2 suggests that larger values of s will favor (and smaller values will disfavor) the use of nonequilibrium trial exchanges.

In summary, we have presented here a method for efficiently incorporating nonequilibrium trials into Hamiltonian exchange. Although it has thus far not been clear whether nonequilibrium methods are capable of achieving greater computational efficiency than traditional equilibrium approaches, the results presented here suggest that under certain circumstances, properly configured HENT does have the potential to achieve at least a modest gain in efficiency. Equation 2 and the accompanying algorithm provide a reasonable way to test whether nonequilibrium trial exchanges are beneficial. To fully validate and gauge the significance of the improvements we have observed, however, further testing involving a variety of systems will be needed.

■ ASSOCIATED CONTENT

S Supporting Information. Included are (1) additional details on hybrid Monte Carlo, (2) a derivation of \widehat{mrtt} , (3) further discussion of the parameter selection algorithm, (4) detailed descriptions of the reaction coordinate and softcore potential, (5) the setup of free energy calculations, (6) technical details related to the simulations performed, (7) technical details related to the Hamiltonian exchange performed, (8) a method for combining conformer data, (9) final setup parameters for simple replica and serial exchange, and (10) a list of nonequilibrium trial exchanges tested during initialization. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: David.Shaw@DEShawResearch.com (D.E.S.); Robert.Dirks@DEShawResearch.com (R.M.D.).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENT

We thank Ron Dror, Michael Eastwood, Albert Pan, and Venkatesh Mysore for help with the manuscript.

■ REFERENCES

- (1) Christen, M.; Van Gunsteren, W. F. On Searching In, Sampling Of, and Dynamically Moving Through Conformational Space of Biomolecular Systems: A Review. *J. Comput. Chem.* **2008**, *29*, 157–166.
- (2) Liwo, A.; Czaplowski, C.; Oldziej, S.; Scheraga, H. A. Computational techniques for efficient conformational sampling of proteins. *Curr. Opin. Struct. Biol.* **2008**, *18* (2), 134–139.
- (3) Leitgeb, M.; Schröder, C.; Boresch, S. Alchemical free energy calculations and multiple conformational substates. *J. Chem. Phys.* **2005**, *122*, 1–15.
- (4) Zwanzig, R. W. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.* **1954**, *22* (8), 1420–1426.
- (5) Marinari, E.; Parisi, G. Simulated Tempering: a New Monte Carlo Scheme. *Europhys. Lett.* **1992**, *19* (6), 451–458.
- (6) Lyubartsev, A. P.; Martsinovski, A. A.; Shevkunov, S. V.; Vorontsov-Velyaminov, P. N. New approach to Monte Carlo calculation of the free energy: Method of expanded ensembles. *J. Chem. Phys.* **1992**, *96*, 1776–1783.
- (7) Hukushima, K.; Nemoto, K. Exchange Monte Carlo Method and Application to Spin Glass Simulations. *J. Phys. Soc. Jpn.* **1996**, *65*, 1604–1608.

- (8) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141–151.
- (9) Mitsutake, A.; Sugita, Y.; Okamoto, Y. Generalized-ensemble algorithms for molecular simulations of biopolymers. *Pept. Sci.* **2001**, *60* (2), 96–123.
- (10) Mitsutake, A.; Okamoto, Y. Multidimensional generalized-ensemble algorithms for complex systems. *J. Chem. Phys.* **2009**, *130* (21), 214105.
- (11) Fukunishi, H.; Watanabe, O.; Takada, S. On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *J. Chem. Phys.* **2002**, *116* (20), 9058–9067.
- (12) Affentranger, R.; Tavernelli, I.; Di Iorio, E. E. A Novel Hamiltonian Replica Exchange MD Protocol to Enhance Protein Conformational Space Sampling. *J. Chem. Theory Comput.* **2006**, *2* (2), 217–228.
- (13) Kannan, S.; Zacharias, M. Enhanced Sampling of Peptide and Protein Conformations Using Replica Exchange Simulations With a Peptide Backbone Biasing-Potential. *Proteins: Struct., Funct., Bioinf.* **2007**, *66*, 697–706.
- (14) Liu, P.; Kim, B.; Friesner, R. A.; Berne, B. J. Replica exchange with solute tempering: A method for sampling biological systems in explicit water. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (39), 13749–13754.
- (15) Huang, X.; Hagen, M.; Kim, B.; Friesner, R. A.; Zhou, R.; Berne, B. J. Replica exchange with solute tempering: Efficiency in large scale systems. *J. Phys. Chem. B* **2007**, *111* (19), 5405–5410.
- (16) Andricioaei, I.; Straub, J. E.; Voter, A. F. Smart Darting Monte Carlo. *J. Chem. Phys.* **2001**, *114* (16), 6994–7000.
- (17) Walter, L.; Weber, M. *Conffump: a fast biomolecular sampling method which drills tunnels through high mountains*; Technical Report 06-26; Konrad-Zuse-Zentrum für Informationstechnik Berlin: Berlin, Germany, 2006, pp 1–9.
- (18) Ytreberg, F. M.; Zuckerman, D. M. Peptide conformational equilibria computed via a single-state shifting protocol. *J. Phys. Chem. B* **2005**, *109* (18), 9096–9103.
- (19) Li, X.; Latour, R. A.; Stuart, S. J. TIGER2: An improved algorithm for temperature intervals with global exchange of replicas. *J. Chem. Phys.* **2009**, *130*, 174106.
- (20) Ballard, A. J.; Jarzynski, C. Replica exchange with nonequilibrium switches. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106* (30), 12224–12229.
- (21) Manousiouthakis, V. I.; Deem, M. W. Strict detailed balance is unnecessary in Monte Carlo simulation. *J. Chem. Phys.* **1999**, *110* (6), 2753–2756.
- (22) Jarzynski, C. Nonequilibrium Equality for Free Energy Differences. *Phys. Rev. Lett.* **1997**, *78* (14), 2690–2693.
- (23) Crooks, G. E. Nonequilibrium Measurements of Free Energy Differences for Microscopically Reversible Markovian Systems. *J. Stat. Phys.* **1998**, *90* (5/6), 1481–1487.
- (24) Brass, A.; Pendleton, B. J.; Chen, Y.; Robson, B. Hybrid Monte Carlo Simulations Theory and Initial Comparison with Molecular Dynamics. *Biopolymers* **1993**, *33*, 1307–1315.
- (25) Predescu, C.; Predescu, M.; Ciobanu, C. V. On the Efficiency of Exchange in Parallel Tempering Monte Carlo Simulations. *J. Phys. Chem. B* **2005**, *109*, 4189–4196.
- (26) Trebst, S.; Troyer, M.; Hansmann, U. H. E. Optimized parallel tempering simulations of proteins. *J. Chem. Phys.* **2006**, *124*, 174903.
- (27) Crooks, G. E. Measuring thermodynamic length. *Phys. Rev. Lett.* **2007**, *99* (10), 100602.
- (28) Shenfeld, D. K.; Xu, H.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. Minimizing thermodynamic length to select intermediate states for free-energy calculations and replica-exchange simulations. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **2009**, *80* (4), 046705.
- (29) Rosta, E.; Hummer, G. Error and efficiency of simulated tempering simulations. *J. Chem. Phys.* **2010**, *132*, 034102.
- (30) Sindhikara, D. J.; Emerson, D. J.; Roitberg, A. E. Exchange often and properly in replica exchange molecular dynamics. *J. Chem. Theory Comput.* **2010**, *6* (9), 2804–2808.
- (31) Abraham, M. J.; Gready, J. E. Ensuring mixing efficiency of replica-exchange molecular dynamics simulations. *J. Chem. Theory Comput.* **2008**, *4* (7), 1119–1128.
- (32) Li, H.; Fajer, M.; Yang, W. Simulated scaling method for localized enhanced sampling and simultaneous “alchemical” free energy simulations: A general method for molecular mechanical, quantum mechanical, and quantum mechanical/molecular mechanical simulations. *J. Chem. Phys.* **2007**, *126*, 024106.
- (33) Muff, S.; Caflisch, A. ETNA: Equilibrium Transitions Network and Arrhenius Equation for Extracting Folding Kinetics from REMD Simulations. *J. Phys. Chem. B* **2009**, *113* (10), 3218–3226.
- (34) Lin, E.; Shell, M. S. Convergence and Heterogeneity in Peptide Folding with Replica Exchange Molecular Dynamics. *J. Chem. Theory Comput.* **2009**, *5* (8), 2062–2073.
- (35) Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Seok, C.; Dill, K. A. Use of the Weighted Histogram Analysis Method for the Analysis of Simulated and Parallel Tempering Simulations. *J. Chem. Theory Comput.* **2007**, *3* (1), 26–41.
- (36) Dreyfus, S. Richard Bellman on the Birth of Dynamic Programming. *Oper. Res.* **2002**, *50* (1), 48–51.
- (37) Guthrie, J. P. A Blind Challenge for Computational Solution Free Energies: Introduction and Overview. *J. Phys. Chem. B* **2009**, *113* (14), 4502–4507.
- (38) Jorgensen, W. L.; Buckner, J. K.; Boudon, S.; Tirado-Rives, J. Efficient computation of absolute free energies of binding by computer simulations. Application to methane dimer in water. *J. Chem. Phys.* **1988**, *89* (6), 3742–3746.
- (39) Fujitani, H.; Tanida, Y.; Ito, M.; Jayachandran, G.; Snow, C. D.; Shirts, M. R.; Sorin, E. J.; Pande, V. S. Direct calculation of the binding free energies of FKBP ligands. *J. Chem. Phys.* **2005**, *123*, 084108.
- (40) Hodel, A.; Rice, L. M.; Simonson, T.; Fox, R. O.; Brünger, A. T. Proline cis-trans isomerization in staphylococcal nuclease: Multi-substrate free energy perturbation calculations. *Protein Sci.* **1995**, *4*, 636–654.
- (41) Woods, C. J.; Essex, J. W. The Development of Replica-Exchange-Based Free-Energy Methods. *J. Phys. Chem. B* **2003**, *107*, 13703–13710.
- (42) Hritz, J.; Oostenbrink, C. Efficient Free Energy Calculations for Compounds with Multiple Stable Conformations Separated by High Energy Barriers. *J. Phys. Chem. B* **2009**, *113*, 12711–12720.
- (43) Chib, S.; Greenberg, E. Understanding the Metropolis-Hastings Algorithm. *Am. Stat.* **1995**, *49* (4), 327–335.
- (44) Shirts, M. R.; Pande, V. S. Solvation free energies of amino acid side chain analogs for common molecular mechanics water models. *J. Chem. Phys.* **2005**, *122*, 134508.
- (45) Andersen, H. C. Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.* **1980**, *72* (4), 2384–2393.
- (46) Bennett, C. H. Efficient Estimation of Free Energy Differences from Monte Carlo Data. *J. Comput. Phys.* **1976**, *22*, 245–268.
- (47) Shirts, M. R.; Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* **2008**, *129* (12), 124105.
- (48) Park, S. Comparison of the serial and parallel algorithms of generalized ensemble simulations: An analytical approach. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **2008**, *77*, 016709.
- (49) Shirts, M. R.; Bair, E.; Hooker, G.; Pande, V. S. Equilibrium Free Energies from Nonequilibrium Measurements Using Maximum-Likelihood Methods. *Phys. Rev. Lett.* **2003**, *91* (14), 140601.
- (50) Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Seok, C.; Dill, K. A. Use of the Weighted Histogram Analysis Method for the Analysis of Simulated and Parallel Tempering Simulations. *J. Chem. Theory Comput.* **2007**, *3* (1), 26–41.
- (51) Bowers, K. J.; Chow, E.; Xu, H.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossváry, I.; Moraes, M. A.; Sacerdoti, F. D.; Salmon, J. K.; Yibing, S.; Shaw, D. E. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. *Proc. ACM/IEEE Conf. Supercomput.*, Tampa, FL, 2006.

(52) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An $N \log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98* (12), 10089–10092.

(53) Lippert, R. A.; Bowers, K. J.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I.; Shaw, D. E. A Common, Avoidable Source of Error in Molecular Dynamics Integrators. *J. Chem. Phys.* **2007**, *126* (4), 046101.

(54) Krautler, V.; van Gunsteren, W. F.; Hunenberger, P. H. A Fast SHAKE Algorithm to Solve Distance Constraint Equations for Small Molecules in Molecular Dynamics. *J. Comput. Chem.* **2001**, *22* (5), 501–508.

Revision of AMBER Torsional Parameters for RNA Improves Free Energy Predictions for Tetramer Duplexes with GC and iGiC Base Pairs

Ilyas Yildirim,[†] Scott D. Kennedy,[‡] Harry A. Stern,[†] James M. Hart,[†] Ryszard Kierzek,[§] and Douglas H. Turner^{*,†}

[†]Department of Chemistry, University of Rochester, Rochester, New York 14627, United States

[‡]Department of Biochemistry and Biophysics, School of Medicine and Dentistry, University of Rochester, Rochester, New York 14642, United States

[§]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 60-714 Poznan, Poland

S Supporting Information

ABSTRACT: All-atom force fields are important for predicting thermodynamic, structural, and dynamic properties of RNA. In this paper, results are reported for thermodynamic integration calculations of free energy differences of duplex formation when CG pairs in the RNA duplexes $r(\text{CCGG})_2$, $r(\text{GGCC})_2$, $r(\text{GCGC})_2$, and $r(\text{CGCG})_2$ are replaced by isocytidine–isoguanosine (iCiG) pairs. Agreement with experiment was improved when ϵ/ζ , α/γ , β , and χ torsional parameters in the AMBER99 force field were revised on the basis of quantum mechanical calculations. The revised force field, AMBER99TOR, brings free energy difference predictions to within 1.3, 1.4, 2.3, and 2.6 kcal/mol at 300 K, respectively, compared to experimental results for the thermodynamic cycles of $\text{CCGG} \rightarrow \text{iCiGiCiG}$, $\text{GGCC} \rightarrow \text{iGiGiCiC}$, $\text{GCGC} \rightarrow \text{iGiCiGiC}$, and $\text{CGCG} \rightarrow \text{iCiGiCiG}$. In contrast, unmodified AMBER99 predictions for $\text{GGCC} \rightarrow \text{iGiGiCiC}$ and $\text{GCGC} \rightarrow \text{iGiCiGiC}$ differ from experiment by 11.7 and 12.6 kcal/mol, respectively. In order to test the dynamic stability of the above duplexes with AMBER99TOR, four individual 50 ns molecular dynamics (MD) simulations in explicit solvent were run. All except $r(\text{CCGG})_2$ retained A-form conformation for $\geq 82\%$ of the time. This is consistent with NMR spectra of $r(\text{iGiGiCiC})_2$, which reveal an A-form conformation. In MD simulations, $r(\text{CCGG})_2$ retained A-form conformation 52% of the time, suggesting that its terminal base pairs may fray. The results indicate that revised backbone parameters improve predictions of RNA properties and that comparisons to measured sequence dependent thermodynamics provide useful benchmarks for testing force fields and computational methods.

1. INTRODUCTION

RNA has a wide variety of biological roles in cells.¹ The genome of some human viruses, such as hepatitis papilloma virus (HPV), human immunodeficiency virus (HIV), smallpox and influenza viruses, is RNA. Messenger RNA (mRNA) carries the code for protein synthesis. Transfer RNAs (tRNA) bring specific amino acids to ribosomes for protein synthesis. Some RNAs, including ribosomal RNA (rRNA), are catalysts.^{2–4} MicroRNAs (miRNA) regulate gene expression.^{5,6} More functions of RNA are still being discovered.

The ability of theoretical and computational approaches to reproduce experimental results provides a test of our understanding of the interactions that shape RNA.⁷ Molecular dynamics (MD) simulations and quantum mechanical (QM) calculations are used to provide insight into biological processes, including folding and dynamics of RNA.^{8–13} The quality of the MD simulations, however, depends on the parametrization of the force fields.

Force fields can be benchmarked against various types of experimental results.⁷ For example, revisions for χ torsional parameters have improved structural predictions of cytidine and uridine,¹⁴ of tetraloop hairpins,¹⁵ and of single-stranded $r(\text{GACC})$.¹⁶ Here, revisions of various torsional parameters are presented and tested against structural and thermodynamic data for duplexes of RNA tetramers.

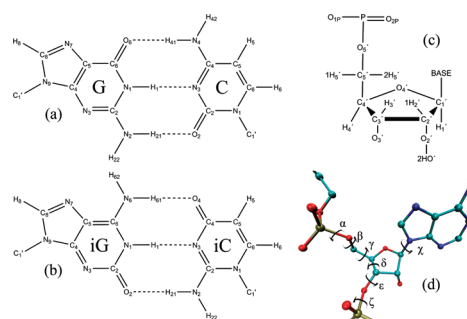


Figure 1. Schematic representations of (a) GC and (b) iGiC base pairs, with atom notations used for (a) guanine (G), cytidine (C), (b) isoguanine (iG), isocytosine (iC), (c) ribose and phosphate, and (d) torsions of nucleic acids.

The unnatural bases isoguanosine (iG) and isocytidine (iC) are similar to the natural bases of guanosine and cytidine except that the amino and carbonyl groups are transposed. They form Watson–Crick-like iGiC base pairs in RNA (Figure 1).^{17,18} UV melting experiments show that the free energies of duplex formation at 300 K of structures with iGiC base pairs are more

Received: August 9, 2011

Published: December 01, 2011

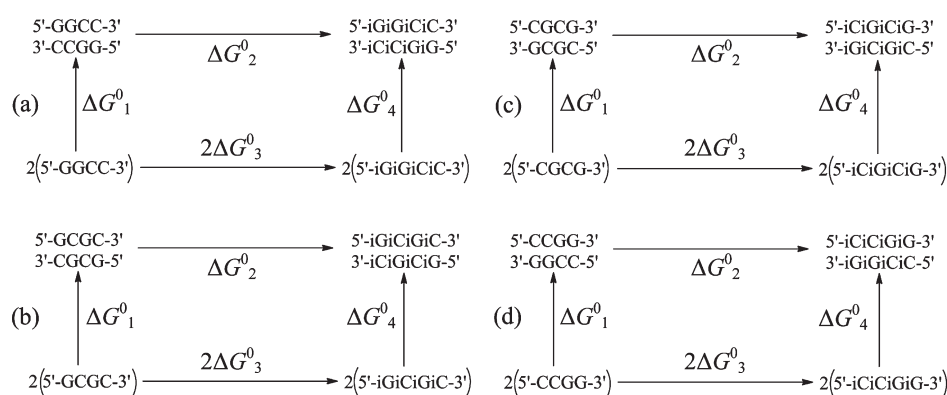


Figure 2. Thermodynamic cycles of (a) GGCC \rightarrow iGiGiCiC, (b) GCGC \rightarrow iCiCiGiG, (c) CGCG \rightarrow iCiCiGiG, and (d) CCGG \rightarrow iCiCiGiG. ΔG^0_2 and ΔG^0_3 represent alchemical transformations of the duplex and single strand, respectively, while ΔG^0_1 and ΔG^0_4 represent duplex formations. Each cycle satisfies the equation of $\Delta G^0_1 + \Delta G^0_2 = 2\Delta G^0_3 + \Delta G^0_4$, where ΔG^0_1 and ΔG^0_4 are experimental values and ΔG^0_2 and ΔG^0_3 are calculated with the TI approach.

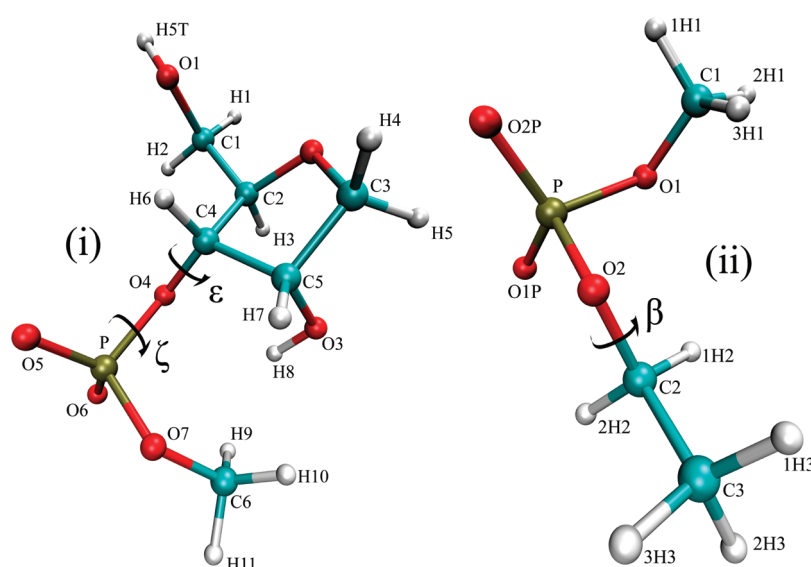


Figure 3. Model systems used to reparameterize torsions ϵ (C2-C4-O4-P), ζ (C4-O4-P-O7), and β (P-O2-C2-C3). 2D and 1D Potential Energy Surface (PES) scans were done to reparameterize ϵ and ζ , and β using model systems (i) and (ii), respectively.

favorable than the structures with GC base pairs.¹⁸ In this paper, $r(iGiGiCiC)_2$ is shown by NMR to have an A-form conformation. Previous NMR and optical melting studies of $r(CCGGp)_2$, where p represents a terminal phosphate, are also consistent with an A-form duplex conformation.^{19–21} These results were generalized for modeling the duplexes of $r(iCiCiGiG)_2$, $r(iCiGiCiG)_2$, $r(iGiCiGiC)_2$, $r(CCGG)_2$, $r(CGCG)_2$, $r(GCGC)_2$, and $r(GGCC)_2$ as A-form structures to allow free energy calculations using the thermodynamic integration (TI)²² approach with thermodynamic cycles shown in Figure 2. The torsional parameters for ϵ , ζ , and β were reparameterized and free energy calculations were made with AMBER99²³ modified with various combinations of parameters for the α/γ ,²⁴ β , ϵ/ζ , and χ ¹⁴ torsions. The version of the AMBER99 force field with revised parameters for all six torsions, which we call AMBER99TOR, improves predictions of differences in experimental free energy changes of duplex formation when iGiC pairs replace GC pairs.

2. METHODS

2.1. Synthesis and Purification of iGiGiCiC. Phosphoramidites for iC and iG were prepared as described previously.¹⁸ The oligoribonucleotide, iGiGiCiC, was synthesized on an Applied Biosystems DNA/RNA synthesizer, using β -cyanoethyl phosphoramidite chemistry.^{25,26} Thin-layer chromatography (TLC) purification of iGiGiCiC was carried out on Merck 60 F254 TLC plates with the mixture 1-propanol/aqueous ammonia/water = 55:35:10 (v/v/v). The details of deprotection and purification of oligoribonucleotides have been described previously.^{27,28}

2.2. NMR. The concentration of the sample was measured with a NanoDrop 2000 Micro-Volume UV-vis spectrophotometer. The NMR sample had 1.65 mM iGiGiCiC in 80 mM NaCl, 10 mM sodium phosphate, and 0.5 mM disodium EDTA at pH 7.0. For spectra in D₂O, two lyophilizations were performed on the sample, reconstituting each time with 99.9% D₂O (Cambridge Isotopes Laboratories), followed by a third lyophilization and reconstitution in 99.990% D₂O (Sigma Aldrich).

Table 1. Conformations Used in 2D ϵ/ζ PES Scan of Model System (i) in Figure 3

conformation	sugar pucker	H5T-O1-C1-C2 (deg)	O1-C1-C2-C4 (deg)	O4-P-O7-C6 (deg)	C3-C5-O3-H8 (deg)
(i)	C2'-endo	174	54	60	-61
(ii)	C2'-endo	174	54	180	-61
(iii)	C2'-endo	174	54	300	-61
(iv)	C3'-endo	174	54	60	-153
(v)	C3'-endo	174	54	180	-153
(vi)	C3'-endo	174	54	300	-153

All spectra were acquired on Varian Inova 500 or 600 MHz NMR spectrometers. Resonances were assigned by standard procedures^{29,30} from NOESY, Watergate NOESY, ¹H-³¹P HETCOR, DQF-COSY, and TOCSY at 0, 20, and 35 °C (see Supporting Information). NOESY spectra were recorded with mixing times (τ_m) of 100, 150, 200, and 400 ms.

2.3. Parametrization. RESP charges for C, G, iC, and iG were calculated as previously described (see Supporting Information).³¹ For C and G, the revised χ torsion parameters of AMBER99 χ were used.¹⁴ The same methodology using Gaussian03³² was applied to reparameterize the χ torsions of iC and iG (see Supporting Information for the parameters). The ϵ and ζ torsions were reparameterized on the basis of 2D potential energy surface (PES) scans on six conformations of model system (i) (Figure 3), defined in Table 1. For each conformation, ϵ and ζ torsions were rotated with increments of 10°, yielding $6 \times (36 \times 36) = 7776$ data points for ϵ/ζ reparameterization. Model system (ii) (Figure 3) was used to reparameterize the β torsion. β torsions were rotated with increments of 10°, yielding 36 data points for β reparameterization. For each conformation in the PES scan, the structures were first optimized with HF/6-31G* level of theory. Then, QM energies were calculated with MP2/6-31G* level of theory. Comparisons between the AMBER99 and revised ϵ/ζ and β torsional parameters are shown in Table 2. For α and γ , torsional parameters of the parmbsc²⁴ force field were used. AMBER99TOR is defined as the AMBER99 force field including all the revised parameters for α/γ ,²⁴ β , ϵ/ζ , and χ ¹⁴ torsions.

2.4. Thermodynamic Cycles. Thermodynamic cycles of CCGG \rightarrow iCiCiGiG, CGCG \rightarrow iCiGiCiG, GCGC \rightarrow iGiCiGiC, and GGCC \rightarrow iGiGiCiC (Figure 2) were used to test free energy calculations with the TI approach. In Figure 2, ΔG^0_1 and ΔG^0_4 are the experimental free energies of duplex formation with GC and iGiC base pairs, respectively. ΔG^0_2 and ΔG^0_3 are the free energies of the alchemical transformations of duplexes and single strands, respectively, from G and C to iG and iC bases. The TI approach with the new mixing rule described previously³¹ was used to calculate ΔG^0_2 and ΔG^0_3 . Each cycle satisfies $\Delta G^0_1 + \Delta G^0_2 = 2\Delta G^0_3 + \Delta G^0_4$.

2.5. Explicit Solvent Simulations. All structures were created with the nucgen module of AMBER9. Structures were solvated with TIP3P³³ water molecules in a truncated octahedral box. In each $S_{G/C} \rightarrow S_{iG/iC}$ alchemical transformation, where $S_{G/C}$ and $S_{iG/iC}$ represent states with G and C and iG or iC bases, respectively, each state had the same number of water molecules (see Supporting Information for the number of water molecules used in each calculation). A total of six and three Na⁺ ions were used to neutralize the duplex and single-stranded RNA systems, respectively. The parameter/topology files for each $S_{G/C} \rightarrow S_{iG/iC}$ transformation were created with the xleap module.²³

Table 2. Comparison of Revised ϵ, ζ, β Torsional Parameters with AMBER99 Counterparts

torsion	n^a	AMBER99 $V_{n,i}^a$	AMBER99TOR $V_{n,i}^a$
ϵ	1	0.000	-1.494
	2	0.000	-0.714
	3	0.383	-0.161
	4	0.000	0.121
ζ	1	0.000	-0.561
	2	1.200	0.575
	3	0.250	-0.997
	4	0.000	-0.078
β	1	0.000	-2.598
	2	0.000	0.011
	3	0.383	-0.322
	4	0.000	-0.082

^aTorsional potential energy in AMBER force field is calculated as $E_{MM,tor}(\phi) = \sum_{i=1}^n V_{n,i} (1 + \cos(n\phi - \gamma))$ where $V_{n,i}$ is the relative potential energy barrier, ϕ is the dihedral angle, γ is the phase shift, and n is the periodicity. For ϵ , ζ , and β torsions, $\gamma = 0$ (see Supporting Information for the modified force field file).

2.5.1. Minimization. Structures were minimized in two steps. For each system, the same protocol was used: (1) RNA structures were 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. The long-range cutoff for nonbonded interactions during minimizations was 8.0 or 10.0 Å. (2) The whole system was minimized without any restraints. Steepest descent minimization of 1000 steps was followed by a conjugate gradient minimization of 1500 steps.

2.5.2. 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 cutoff of 8.0 or 10.0 Å was used. SHAKE³⁴ was turned on for bonds involving hydrogen atoms, except for the amino hydrogen and dummy atoms. Temperature was raised from 0 to 300 K in 20 ps. Langevin dynamics with a collision frequency of 1 ps⁻¹ was used. Ten thousand MD steps were run with a 2 fs time step, yielding a total of 20 ps of MD. (2) The same conditions as above were chosen, except that no restraints on the structures and constant pressure dynamics were used. Reference pressure was set to 1 atm with a pressure relaxation time of 2 ps. A total of 100 ps of MD was run with a 2 fs time step. The final restart file was used as the starting structure

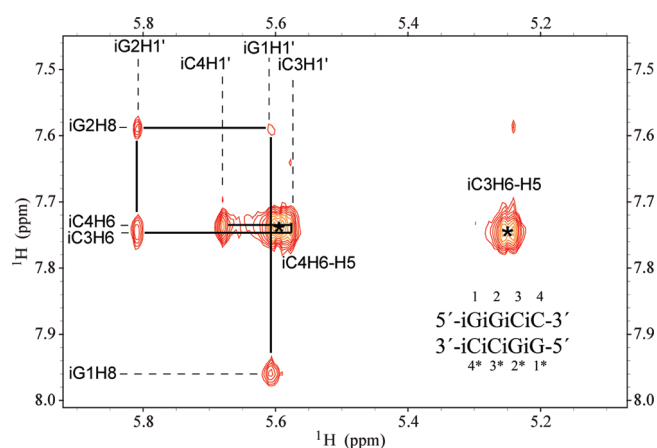


Figure 4. NOESY walk of $(iGiGiCiC)_2$ from 200 ms mixing time NOESY experiment at 20 °C. Because the sequence is symmetric, the cross-peaks from each stand overlap. Residue numberings are shown in the bottom right corner.

for the λ simulations. In the AMBER99 force field simulations, constant volume dynamics were used.

2.5.3. λ Simulations. Nineteen λ values were used; $\lambda = 0.05$ to $\lambda = 0.95$, with an increment of 0.05. The new mixing rule of TI approach was used in all λ simulations.³¹ For each λ simulation, the last structure of pressure regulation was taken as the initial structure. The production run was similar to the second step of the pressure equilibration described above. Duplex and single-strand MD simulations, respectively, were run for 2 and 3 ns with 1 fs time steps.

2.5.4. Restrained λ Simulations. Additional MD simulations used dihedral restraints to restrict the sampling space to A-form conformations (see Supporting Information). A total of 1 ns of MD was run with a 1 fs time step for both the duplex and single-strand simulations. Further calculations used positional restraints with weight of 10 kcal/mol Å² on the backbone heavy atoms in single-strand simulations.

2.5.5. Dynamic Stabilities of RNA Duplexes with AMBER99-TOR Force Field. In order to analyze the dynamic stability of each duplex and single-strand with the AMBER99TOR force field, MD simulations in explicit solvent were run. Systems were prepared similar to part 2.5. Six and three Na⁺ ions were used to neutralize duplex and single-strand systems, respectively. Duplex and single-strand systems were solvated with 1786 and 1345 TIP3P³³ water molecules, respectively, in a truncated octahedral box. The systems were minimized and pressure regulated as described above. Each production run included 50 ns of MD with 1 fs time step at 300 K. Trajectory files were written at each 500 fs time step. Four individual simulations were run for each system, yielding a total of 200 ns of MD.

2.6. Analysis. Free Energy Calculations Using TI Approach. The first 250 ps of each λ simulation were omitted from the calculations. For each λ simulation, $\langle \partial E / \partial \lambda \rangle_\lambda$ was calculated. The trapezoidal rule was used to numerically integrate $\langle \partial E / \partial \lambda \rangle_\lambda$ vs λ curves to get ΔG^0 . Multiple transformations were done to calculate the means and standard deviations (see Supporting Information).

Stability Analysis of Duplex Simulations with AMBER99TOR Force Field. The combined 200 ns of MD simulations were analyzed to test the dynamic stability of each duplex and single strand (see Supporting Information). All the trajectory data were

rmsd fitted to the initial A-form starting structure. For each simulation, the ptraj module of AMBER 9³³ was used to calculate the percentage of structures having an all-atom rmsd less than 1.5 and 3.0 Å. Qualitatively, A-form and A-form-like structures are defined as conformations with rmsd less than 1.5 Å and 3.0 Å, respectively. Total overlap area of the stacked base pairs were calculated with 3DNA³⁵ using snapshots extracted from the trajectories at intervals of 0.5 ns.

3. RESULTS

3.1. Conformation of $r(iGiGiCiC)_2$. Because the electronic structure of iCiG base pairs differs from CG,⁷ NMR was used to test the expectation that $r(iGiGiCiC)_2$ has an A-form conformation. Figure 4 shows the NOESY walk region of $(iGiGiCiC)_2$ from a 200 ms mixing time NOESY experiment in D₂O at 20 °C. NMR distance limits were extracted from NOESY spectra at 20 and 35 °C with 200 ms mixing time using intranucleotide H1'/H2' cross-peaks as reference NOEs (see Supporting Information).

At 1.65 mM iGiGiCiC, there is an iG1H1'/iC4H2' cross-peak. This cross-peak disappeared when the iGiGiCiC concentration was diluted to 0.17 mM (see Supporting Information) and is due to coaxial stacking of duplexes. The rest of the spectrum was essentially unchanged at the lower concentration.

The NMR spectra of iGiGiCiC are consistent with an A-form duplex conformation. All the residues prefer C3'-endo sugar pucker as evidenced by ³J_{H1'-H2'} couplings of less than 2 Hz as estimated from peak splittings. Intranucleotide iGH8/H1' and iCH6/H1' cross-peaks have volumes indicating anti conformations. A Watergate NOESY spectrum at 0 °C with 150 ms mixing time showed iG1H61-iC3H5 and iG2H1-iC4H1' cross-peaks consistent with a duplex structure in which these peaks are actually iG1H61-iC3'H5, iG1'H61-iC3H5, iG2H1-iC4'H1', and iG2'H1-iC4H1', where an asterisk represents the opposite RNA strand (see Supporting Information). The chemical shifts of the iG imino protons (Supporting Information) are consistent with hydrogen bonding, as observed for other iCiG pairs.^{17,18} Separate resonances are seen for the two protons of the iG amino groups with one of the shifts consistent with hydrogen bonding.^{17,18} Another expectation for a "Watson-Crick" iGiC pair is a cross-peak from iGH1 to both protons of an iC amino group, and iG2 shows such cross-peaks to two broad resonances. A HETCOR spectrum showed phosphorus shifts within 0.3 ppm, implying regular A-form conformation (see Supporting Information). The HETCOR spectrum also showed strong (n)P-(n-1)H3' and weak (n)P-(n)H5'/5'' scalar coupling typical of A-form ϵ and β conformations and weak H4'-H5'/5'' scalar coupling consistent with A-form γ conformation. Distances measured for the nucgen model of $(iGiGiCiC)_2$ are consistent with the distance limits calculated from NOEs, with the exception of a 0.15 Å difference for iG2'H3'-iG2'H8 (see Supporting Information). On the basis of the NMR spectra for $r(iGiGiCiC)_2$, A-form conformations were also modeled for $r(iCiCiGiG)_2$, $r(iCiGiCiG)_2$, and $r(iGiCiGiC)_2$.

3.2. Comparisons of Molecular Mechanics (MM) to QM Energies before and after ϵ/ζ and β Reparameterizations. Model systems (i) and (ii) (Figure 3) were used to reparameterize the ϵ/ζ and β torsional parameters, respectively. A force field with the new ϵ/ζ parameters is called AMBER99EZ. Figure 5 shows comparisons of the 2D potential energy surfaces approximated by AMBER99EZ (see Supporting Information for the definitions) and AMBER99 force fields with the QM potential energy surfaces for six conformations (see Supporting Information).

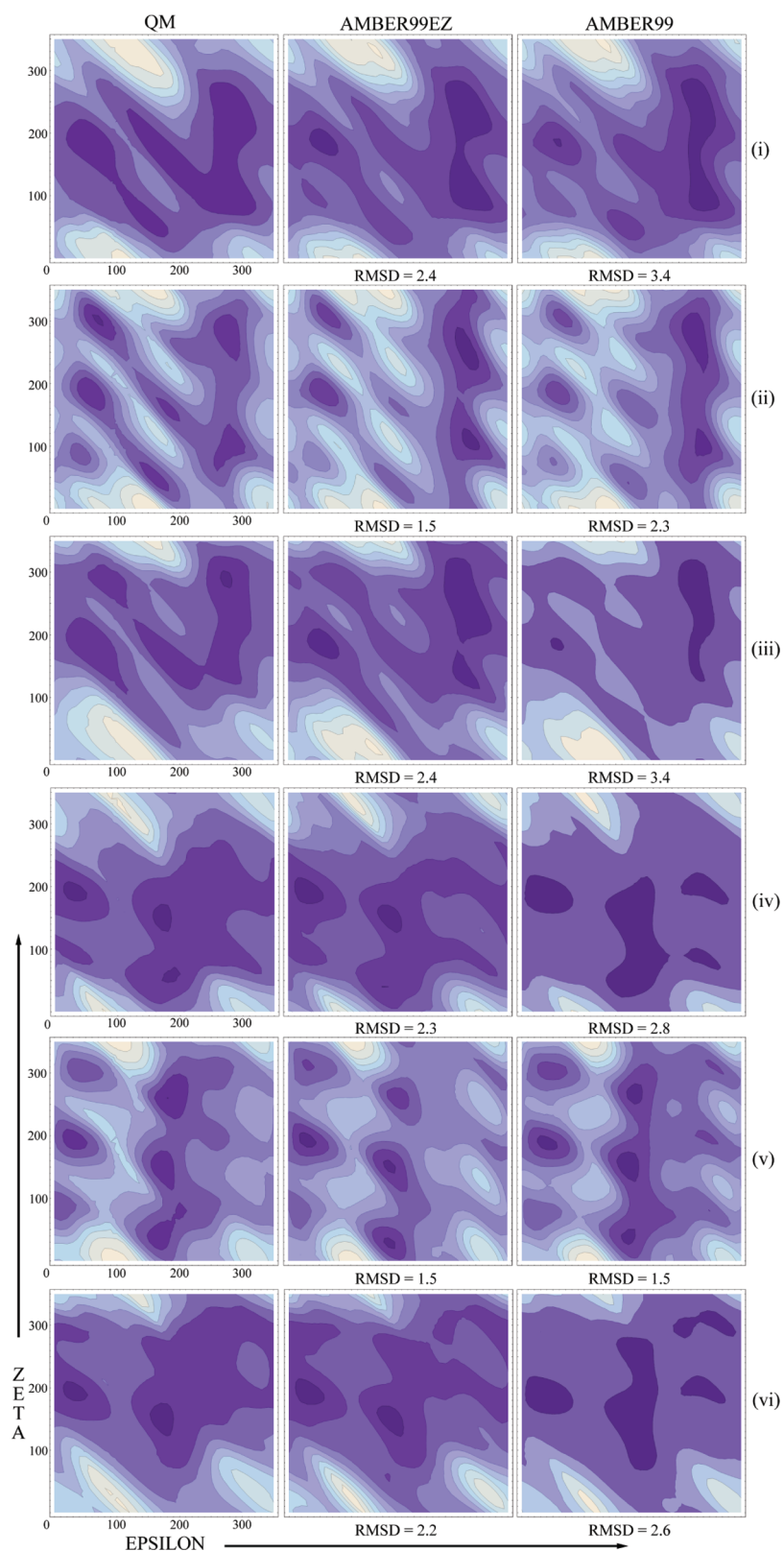


Figure 5. Two-dimensional PES scans of ϵ (x -axis) and ζ (y -axis) for six conformations described in Table 1. QM, AMBER99, and AMBER99EZ stand for PES scans of the conformations (i–vi) using quantum mechanics, AMBER99, and reparameterized ϵ/ζ torsional set of AMBER99 force field, respectively. rmsd values (kcal/mol) under each PES scan are with respect to QM. The darker the violet color, the lower the energy value.

Figure 6 shows the 1D potential energy surfaces for model system (ii) of Figure 3 as calculated by QM, AMBER99 with revised

β torsional parameters, and AMBER99. It is clear from Figures 5 and 6 that revisions of the torsional parameters improve the

approximations of the QM potential energy surfaces for these model systems. The revision of the AMBER99 force field uses four cosine terms to describe each of the torsional energy profiles of ϵ/ζ and β , while the original AMBER99 force field uses one cosine term each for ϵ and β , and two cosine terms for ζ to describe the torsional energy profiles (Table 2). Increasing the number of cosine terms evidently improves the predictions of QM PES profiles.

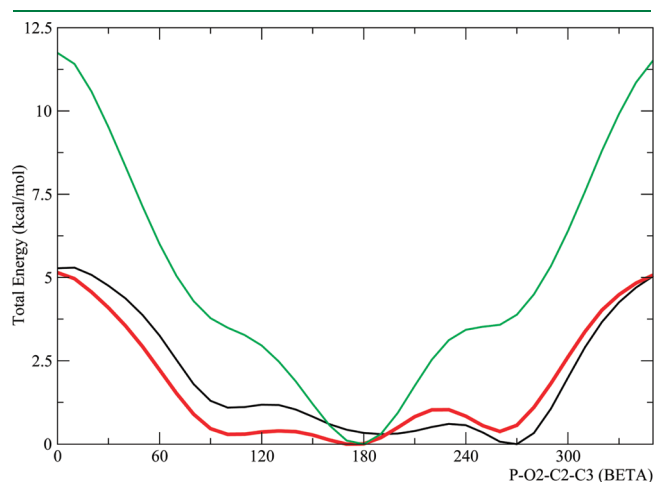


Figure 6. Potential energy (kcal/mol) vs β dihedral angle (P–O2–C2–C3) of model system (ii) (Figure 3) with QM (black), AMBER99 with revised β torsion parameter (red), and AMBER99 (green). For visualization purposes, minimum energies are set to zero.

3.3. Comparisons between Measured and Calculated Changes in Free Energies of Duplex Formation by CG and iCiG Sequences to Benchmark the Effects of Revising Torsional Parameters. As illustrated in Figure 2, experimental measurements^{7,18,19,28,36,37} provide values for the free energy changes, ΔG^0_1 and ΔG^0_4 , of formation of duplexes with CG and iCiG pairs, respectively. The values of ΔG^0_2 and ΔG^0_3 for the alchemical transitions in Figure 2 can be calculated with the TI approach,^{22,31} where ΔG^0_3 represents the change for one single strand morphing from C and G to iC and iG nucleotides. From the thermodynamic cycles in Figure 2, $\Delta G^0_4 - \Delta G^0_1 = \Delta G^0_2 - 2\Delta G^0_3$. Thus comparisons of experimental values for $\Delta G^0_4 - \Delta G^0_1$ with calculated values for $\Delta G^0_2 - 2\Delta G^0_3$ provide benchmarks for the effects of revising force field parameters. This is a rather stringent test because the individual values for ΔG^0_2 and $2\Delta G^0_3$ are on the order of 200 kcal/mol (see Supporting Information), while the experimental values for $\Delta G^0_4 - \Delta G^0_1$ are on the order of a few kcal/mol (Table 3).

Table 3 shows results from unrestrained and restrained simulations with several combinations of torsion parameters. Initial results for unrestrained simulations with AMBER99 on the GCGC \rightarrow iGiCiGiC and GGCC \rightarrow iGiGiCiC cycles gave values for $\Delta G^0_2 - 2\Delta G^0_3$ of 8.7 and 10.5 kcal/mol, respectively, whereas the experimental values for $\Delta G^0_4 - \Delta G^0_1$ are -3.0 and -2.1 kcal/mol, respectively. In contrast, AMBER99 simulations with the backbone torsions restrained to A-form gave $\Delta G^0_2 - 2\Delta G^0_3$ values of -1.4 and -1.5 kcal/mol, respectively. When the positions of backbone heavy atoms were restrained, the values were -3.0 and -2.4 kcal/mol, respectively, close to the experimental values. The lack of agreement for unrestrained AMBER99

Table 3. Free Energy Results (kcal/mol at 300 K) of Unrestrained and Restrained Simulations for the Thermodynamic Cycles of GCGC \rightarrow iGiCiGiC, GGCC \rightarrow iGiGiCiC, CGCG \rightarrow iCiGiCiG, and CCGG \rightarrow iCiCiGiG Using TI Approach with Revised Torsions for AMBER99 Force Field

thermodynamic cycle	$\Delta G^0_2 - 2\Delta G^0_3$				$\Delta G^0_4 - \Delta G^0_1$ ^a
	none ^b	χ ^b	$\chi\alpha\gamma$ ^b	$\chi\alpha\gamma\epsilon\zeta$ ^b	
	revised torsions				
				$\chi\alpha\gamma\epsilon\zeta\beta$ ^b	
	Unrestrained Simulations				
GCGC \rightarrow iGiCiGiC	8.7	-1.2 ± 0.4	-0.7 ± 1.2	-1.3 ± 1.0	-3.0 ± 0.4
GGCC \rightarrow iGiGiCiC	10.5	-1.3 ± 0.6	0.1 ± 2.4	-6.2 ± 2.4	-2.1 ± 0.4
CGCG \rightarrow iCiGiCiG	–	-0.8 ± 1.6	0.0 ± 0.4	0.2 ± 1.9	-2.2 ± 0.4
CCGG \rightarrow iCiCiGiG	–	4.5 ± 1.6	1.7 ± 1.0	2.1 ± 3.4	-0.4 ± 0.3
rmsd ^c	12.2	2.7	2.2	2.8	2.0
	Restrained Simulations ^d				
GCGC \rightarrow iGiCiGiC	-1.4 (-3.0)	–	–	-1.8 ± 0.4	-3.0 ± 0.4
GGCC \rightarrow iGiGiCiC	-1.5 (-2.4)	–	–	-1.4 ± 0.2	-2.1 ± 0.4
CGCG \rightarrow iCiGiCiG	-1.1	–	–	-1.2 ± 0.4	-2.2 ± 0.4
CCGG \rightarrow iCiCiGiG	-0.4	–	–	-0.5 ± 0.4	-0.4 ± 0.3
rmsd ^c	1.0			0.9	1.0

^a These values are experimental results at 300 K.¹⁸ Error limits assume $\pm 4\%$ error for each ΔG^0 .²⁸ ^b none = AMBER99, χ = AMBER99 χ ,¹⁴ $\chi\alpha\gamma$ = AMBER99 χ ¹⁴ + parmbsc,²⁴ $\chi\alpha\gamma\epsilon\zeta$ = AMBER99 χ + parmbsc + AMBER99EZ, $\chi\alpha\gamma\epsilon\zeta\beta$ = AMBER99TOR (see Supporting Information for the definitions of these force fields). ^c rmsd = $(\frac{1}{4}\sum_{i=1}^4 (\Delta G^0_{\text{calculated},i} - \Delta G^0_{\text{measured},i})^2)^{1/2}$ where $\Delta G^0_{\text{calculated}} = \Delta G^0_2 - 2\Delta G^0_3$, $\Delta G^0_{\text{measured}} = \Delta G^0_4 - \Delta G^0_1$, and i stands for results of each thermodynamic cycle. ^d Values not in parentheses are for simulations with dihedral restraints (see Supporting Information). Values in parentheses are simulations with positional restraints. Restraint weight of 10 kcal/mol \AA^2 was applied to backbone heavy atoms in single-strand MD simulations to force them to sample around A-form conformations. No restraints were used in duplex simulations for these calculations.

Table 4. All-Atom RMSD (Å) Results and Total Overlap Area of the Stacked Base Pairs for Duplex Simulations of (CCGG)₂, (CGCG)₂, (GCGC)₂, (GGCC)₂, (iCiCiGiG)₂, (iGiGiCiG)₂, (iGiGiCiC)₂, and (iGiGiCiC)₂ with AMBER99TOR^a

duplex	≤1.5 (%)	≤3.0 (%)	overlap area ^b (Å ²)	duplex	≤1.5 (%)	≤3.0 (%)	overlap area ^b (Å ²)	$\Delta G_4^0 - \Delta G_1^0$ ^c
(GCGC) ₂	39 ± 29	98 ± 5	10.6 ± 0.3	(iGiGiCiC) ₂	64 ± 18	100 ± 0	10.4 ± 0.1	-3.0
(GGCC) ₂	47 ± 16	100 ± 0	7.1 ± 0.1	(iGiGiCiC) ₂	85 ± 6	100 ± 0	6.7 ± 0.0	-2.1
(CGCG) ₂	28 ± 15	95 ± 6	7.0 ± 0.0	(iCiCiGiG) ₂	22 ± 12	82 ± 28	6.5 ± 0.1	-2.2
(CCGG) ₂	15 ± 6	51 ± 33	4.5 ± 0.0	(iCiCiGiG) ₂	29 ± 16	95 ± 10	4.8 ± 0.1	-0.4

^aFor each duplex, four individual MD simulations of 50 ns were run at 300 K, yielding a total of 200 ns. Structures were saved every 0.5 ps for rmsd analysis. ^bTotal overlap area of the stacked base pairs excluding exocyclic groups was calculated with 3DNA³⁵ using all the snapshots extracted from the trajectories at intervals of 0.5 ns. ^c ΔG_1^0 and ΔG_4^0 are duplex formation free energies of structures with GC and iGiC base pairs, respectively, at 300 K.¹⁸

calculations suggests poor sampling of the conformational space. These results suggested that revisions of torsional parameters could improve agreement between calculations and experiments.

To test the effects of adding revised torsional parameters, unrestrained TI calculations were done on all four systems shown in Figure 2. Specifically, unrestrained calculations were done with AMBER99 χ ¹⁴, AMBER99 χ with α/γ parameters from parmbsc²⁴ ($\chi\alpha\gamma$), and with further revision of parameters for ϵ/ζ ($\chi\alpha\gamma\epsilon\zeta$), or ϵ/ζ and β ($\chi\alpha\gamma\epsilon\zeta\beta$, AMBER99TOR) as developed here (Table 3).

As shown in Table 3, all revisions improved agreement between predictions and experiments relative to AMBER99 calculations for GCGC \rightarrow iGiGiCiC and GGCC \rightarrow iGiGiCiC. Revision of χ parameters provided a large improvement of 10–12 kcal/mol at 300 K. Revisions for other dihedral parameters were tested against experimental results for all four thermodynamic cycles shown in Figure 2. AMBER99TOR gave the best rmsd of 2.0 kcal/mol between predictions and experiment, but AMBER99 χ mixed with α/γ parameters taken from parmbsc was similar with an rmsd of 2.2 kcal/mol (Table 3).

Relatively large error limits of 2.4 kcal/mol were found for AMBER99TOR calculations of GGCC \rightarrow iGiGiCiC and CCGG \rightarrow iCiCiGiG (Table 3). Nevertheless, the calculated values of $\Delta G_2^0 - 2\Delta G_3^0$ for these transformations are within 1.4 kcal/mol of the experimental $\Delta G_4^0 - \Delta G_1^0$, well within experimental error. In contrast, the calculated values for GCGC \rightarrow iGiGiCiC and CGCG \rightarrow iCiCiGiG have error limits of 1.0 kcal/mol, but values of $\Delta G_2^0 - 2\Delta G_3^0$ differ from $\Delta G_4^0 - \Delta G_1^0$ by 2.3 and 2.6 kcal/mol, respectively. These comparisons suggest a difference in the behavior of sequences with 5'GG/3'CC and 5'iGiG/3'iCiC nearest neighbors and those without adjacent G's. Root-mean-square deviation analysis of each λ simulation with AMBER99TOR showed that all the duplex transformations (corresponding to ΔG_2^0 in Figure 2) sample pure A-form conformations over 80% of the time while the single-strand transformations (corresponding to ΔG_3^0 in Figure 2) behave differently for sequences with 5'GG/3'CC nearest neighbors (see Supporting Information). The single-strand transformations of CGCG \rightarrow iCiCiGiG and GCGC \rightarrow iGiGiCiC sample A-form conformations 47% of the time on average, whereas the single-strand transformations of CCGG \rightarrow iCiCiGiG and GGCC \rightarrow iGiGiCiC sample A-form only 21% of the time on average (see Supporting Information). As a result, errors for the thermodynamic cycles of CCGG \rightarrow iCiCiGiG and GGCC \rightarrow iGiGiCiC are 2.4 kcal/mol while they are 1.0 kcal/mol for the cycles of CGCG \rightarrow iCiCiGiG and GCGC \rightarrow iGiGiCiC (Table 3). The more the single-strands sample A-form conformations in a thermodynamic cycle, the lower the error.

A study of the ability of AMBER99 to predict experimentally observed¹⁷ relative populations of sheared and imino hydrogen bonded GA pairs in RNA duplexes found that the best agreement required restraining backbones to be similar to those known for the NMR structures.³¹ As described above, restraining backbones to A-form conformations dramatically increased agreement of AMBER99 calculations with experiment. To test the effects of dihedral restraints on revised versions of AMBER99, calculations were done with AMBER99TOR and AMBER99 χ with α/γ and ϵ/ζ revisions (Table 3). In both cases, agreement with experiment was improved with RMSDs of 1.0 and 0.9 kcal/mol, respectively, and error limits were reduced to 0.4 or fewer kcal/mol. Even with dihedral restraints, however, experimental and calculated values sometimes differ beyond error limits. The results suggest that the force field can be refined further. It is encouraging, however, that the CCGG \rightarrow iCiCiGiG transformation is predicted to have the smallest $\Delta G_2^0 - 2\Delta G_3^0$, which agrees with experiment.

3.4. Predicted Dynamic Stabilities with AMBER99TOR Force Field. The predicted dynamic stability of duplexes provides another test of force fields. For each duplex, four individual MD simulations of 50 ns each were run with the AMBER99TOR force field and then combined for an rmsd analysis. The percentage of structures having all-atom rmsd less than 1.5 or 3.0 Å were calculated (Table 4). rmsd \leq 1.5 Å and rmsd \leq 3.0 Å represent, respectively, essentially the starting A-form conformation and A-form-like conformations.

Table 4 shows the percentage of structures in A-form and A-form-like conformations for the duplexes over 200 ns of unrestrained MD with the AMBER99TOR force field. All duplexes except r(CCGG)₂ spent more than four-fifths of time in A-form and A-form-like conformations. The results suggest that r(CCGG)₂ may have unusual dynamics for its terminal base pairs, e.g., "fraying". The MD simulations for the other sequences indicate that A-form-like conformations are essentially stable for at least 50 ns with the AMBER99TOR force field. MD simulations of single-strands show that A-form conformations are sampled rarely while even A-form-like conformations are sampled less than 60% of time (see Supporting Information).

3.5. Comparison between Predicted Values of β and Those Observed in Crystal Structures. QM calculations on model system (ii) in Figure 3 predict a shallow dependence of energy on the β dihedral angle (Figure 6). Crystal structures of RNA, however, show a strong preference for $\beta \sim 180^\circ$,^{38,39} as does the AMBER99 potential (Figure 6). Histogram analysis of unrestrained 50 ns MD simulations with AMBER99TOR that correspond to a total of 3.2 μ s simulation time shows two populations preferred by β (Figure 7). The dominant region has β around 180° , while the minor region is around 80° . While low

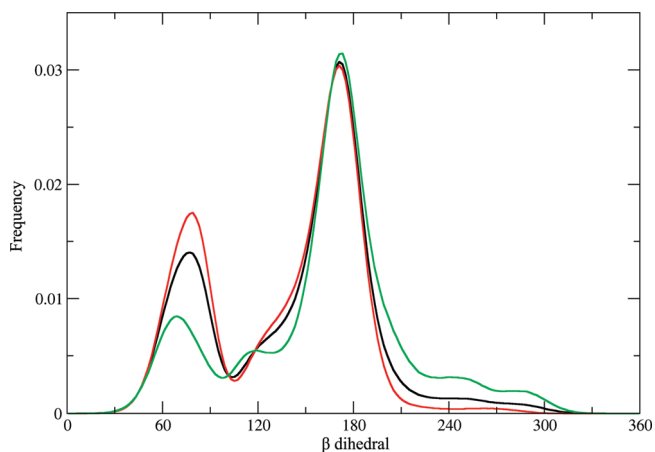


Figure 7. Population distribution analysis for β torsion of MD simulations with AMBER99TOR. Black, red, and green curves represent results including all, duplex, and single-strand MD simulations, respectively. Terminal β torsions (free 5'OH) were omitted from calculations.

values of β are rarely seen in RNA crystals, they are seen in RNA S-motifs.³⁸

The β dihedral is coupled with α and γ torsions. Cluster analysis of the unrestrained MD simulations showed that there are three (α, β, γ) populations; 66% in ($300^\circ, 180^\circ, 60^\circ$), 17% in ($180^\circ, 80^\circ, 180^\circ$), and 13% in ($300^\circ, 80^\circ, 180^\circ$). Even though there are three regions preferred by (α, β, γ), 3D structures created by these combinations are similar (see Supporting Information). Crystal structures analyzed by others^{38,39} are much bigger systems compared to tetramers discussed here. This might explain why we see 30% of structures with low β values. It is also possible that this low β region might have importance for backbone dynamics such as in base unstacking and base pair opening.

3.6. Comparison between Sequence Dependence of Free Energy Differences and Overlap of Bases. There is a parallelism between $\Delta G_4^0 - \Delta G_1^0$ values for the cycles shown in Figure 2 and total overlap areas of stacked base pairs for the duplexes studied (Table 4). Total overlap areas of stacked base pairs for (GCGC)₂ and (iGiCiGiC)₂, (GGCC)₂ and (iGiCiGiC)₂, (CGCG)₂ and (iCiGiCiG)₂, (CCGG)₂ and (iCiCiGiGi)₂ are around 10.5, 6.9, 6.8, and 4.7 Å², respectively, in parallel with experimental $\Delta G_4^0 - \Delta G_1^0$ results for the corresponding thermodynamic cycles (Table 4). This suggests that the thermodynamic differences between duplexes with CG and iCiG pairs may be primarily due to differences in stacking interactions, which result from different electron distributions in the ring systems.

4. DISCUSSION

Force fields for proteins have proven to be extremely useful for providing insight into protein folding, function, and design.^{40–43} Much less effort, however, has been applied to development and testing of force fields for RNA. The emerging recognition that RNA has many different cellular functions^{1–6} and that many RNAs are potential therapeutic targets^{27,44–47} increases the importance of force fields for RNA.

A key aspect of RNA is base orientation with respect to sugar. This orientation is controlled by the χ torsions, which define whether a nucleotide is in syn or anti conformation. Revision of χ torsion parameters to give the AMBER99 χ force field improves structural and thermodynamic predictions for cytidine and uridine¹⁴

and structural predictions for tetraloop hairpins.¹⁵ For example, AMBER99 prefers syn base orientation for pyrimidines, while AMBER99 χ prefers anti base orientation.¹⁴ Additionally, AMBER99 prefers either syn or high-anti base orientations for purines, while AMBER99 χ prefers syn or anti base orientations.¹⁴ Structural analysis of single-stranded r(GACC) with NMR showed that it prefers A-form-like conformations. AMBER99, however, rapidly generates random coil conformations for r(GACC) while AMBER99 χ prefers A-form-like conformations for most of the first 700 ns of an unrestrained MD simulation.¹⁶ After 700 ns, however, a stable conformation and random coil ensemble were generated that are inconsistent with NMR spectra.

A different reparameterization of χ ⁴⁸ has been tested along with AMBER99 χ for ability to maintain known RNA structures during unrestrained MD simulations.^{15,48} Both revisions performed better than AMBER99 and similarly to each other even though there were many differences in the details of methods used for parametrization.^{14,48} Thus, there is consensus that parameters for χ are important for accurately modeling RNA.

Here, various versions of AMBER99 are developed and benchmarked against measured differences in the free energies of duplex formation by tetramers with GC or iGiC base pairs.^{18,19,36,37} Comparisons between measurements and computations are based on the thermodynamic cycles shown in Figure 2, and the results are listed in Table 3.

Relative to AMBER99, AMBER99 χ improves agreement between experiments and predictions for the GCGC \rightarrow iGiCiGiC and GGCC \rightarrow iGiCiGiC cycles by about 11 kcal/mol at 300 K, corresponding to a 10⁸ improvement in prediction of relative equilibrium constants. When AMBER99 χ is tested against further refined force fields, the best agreement with experiment for all four cycles shown in Figure 2 is found with AMBER99-TOR, which includes α/γ parameters from the parmbsc force field,²⁴ along with new parameters developed here for ϵ/ζ and β (Tables 2 and 3). These additional parameters improve the rmsd comparison by 0.7 kcal/mol at 300 K, corresponding to a 3-fold improvement in prediction of relative equilibrium constants. The largest improvement, however, is 3.6 kcal/mol for CCGG \rightarrow iCiCiGiG at 300 K, corresponding to a 400-fold improvement in relative equilibrium constants.

TI calculations without restraints or with dihedral restraints did not predict the magnitudes of all the experimental results within error limits (Table 3). This implies that approximations can be improved. The free energy differences for formation of duplexes from single strands depend on many interactions, including stacking, hydrogen bonding, and solvent interactions in both single strands and duplexes. For example, treatment of van der Waals interactions may need revision for RNA force fields to better predict experimental results. Free energy difference calculations provide useful benchmarks for testing such force field revisions.

The ability of force fields to maintain known 3D structures is another test. Here, NMR spectra (Figure 4 and Supporting Information) show that (iGiCiGiC)₂ is an A-form duplex. All the other duplexes are also expected to be A-form except for occasional fraying of terminal base pairs. As shown in Table 4, results from four unrestrained 50 ns MD simulations for each of the eight duplexes studied are consistent with this expectation. It is encouraging that AMBER99TOR appears to provide reasonable results for both free energy calculations and dynamics of tetramer duplexes containing either GC or iGiC pairs.

The results presented here show that AMBER99 χ and AMBER99TOR improve predictions of the sequence dependence of thermodynamics for several tetramer duplexes and that MD simulations with AMBER99TOR usually retain A-form like structure for at least 50 ns. The revised force fields have not been tested on larger RNAs. It would be surprising, however, if they did not also work well for larger RNAs where the accessible folding space is more limited by volume exclusion.

5. CONCLUSION

Differences in stabilities of short RNA duplexes provide tests of computational methods and force fields. The tests are especially stringent because the calculations include single strands, which have conformational flexibility without much restriction from volume exclusion or hydrogen bonding. Comparisons between measured and predicted stabilities of tetramer duplexes with either GC or iGiC base pairs reveal that reparameterization of torsions can improve agreement between experiment and computations by roughly 10 kcal/mol at 300 K, corresponding to an improvement of about 10^7 in relative equilibrium constant. Most of the improvement relies on new parameters for the χ torsion. The new parameters also largely retain A-form like structures in 50 ns long MD simulations. The revised parameters should improve computations of properties for RNA loops. Loops are often important for function and have weaker interactions and more dynamics than stems. The results also indicate, however, that computations can be further improved.

■ ASSOCIATED CONTENT

S Supporting Information. Description of RESP charge calculations of model systems and definition of each torsional revision; atom names/types/charges of model systems, and nucleotides G, iG, C, and iC; modified force field files for revised torsional parameters of χ , α/γ , ϵ/ζ , β , and nucleotides iC and iG; TI results of all the thermodynamic cycles with different revisions of the torsions (restrained/unrestrained calculations); dihedral restraints used in restrained TI calculations; rmsd analysis of each λ simulation of all the thermodynamic cycles with AMBER99TOR; number of water molecules used to solvate the duplex and single-strand structures; NMR distance, Watson–Crick base pairing, dihedral, and chirality restraints used in simulated annealing of iGiGiCiC; NMR chemical shift assignments of $r(iGiGiCiC)_2$; NOESY walk, ^1H – ^{31}P HETCOR and NOESY spectra of iGiGiCiC; population distribution analysis of torsions; overlap of stable three (α , β , γ) combinations seen in AMBER99TOR MD simulations. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: (585) 275-3207. Fax: (585) 276-0205. Email: turner@chem.rochester.edu

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENT

We are grateful to the Center for Research Computing at the University of Rochester for providing the necessary computing

systems (BlueHive cluster) and personnel to enable the research presented in this manuscript, and Jason D. Tubbs for his help with the NanoDrop 2000 UV–vis spectrophotometer. Dr. Ilyas Yildirim thanks the Faculty of Science, Akdeniz University, Antalya, Turkey, for providing a working environment while he was in Turkey. We also thank the AMBER community and mailing list for all their help and support. This work was supported by NIH grant GM22939 (D.H.T.).

■ ABBREVIATIONS USED

C, cytidine; G, guanosine; iC, isocytidine; G, isoguanosine; QM, quantum mechanics; MM, molecular mechanics; MD, molecular dynamics; TI approach, Thermodynamic Integration approach

■ REFERENCES

- (1) Atkins, J. F.; Gesteland, R. F.; Cech, T. R. *RNA Worlds: From Life's Origins to Diversity in Gene Regulation*; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, 2010.
- (2) Guerrier-Takada, C.; Gardiner, K.; Marsh, T.; Pace, N.; Altman, S. *Cell* **1983**, *35*, 849.
- (3) Kruger, K.; Grabowski, P. J.; Zaug, A. J.; Sands, J.; Gottschling, D. E.; Cech, T. R. *Cell* **1982**, *31*, 147.
- (4) Nissen, P.; Hansen, J.; Ban, N.; Moore, P. B.; Steitz, T. A. *Science* **2000**, *289*, 920.
- (5) Lee, R. C.; Feinbaum, R. L.; Ambros, V. *Cell* **1993**, *75*, 843.
- (6) Ruvkun, G. *Science* **2001**, *294*, 797.
- (7) Yildirim, I.; Turner, D. H. *Biochemistry* **2005**, *44*, 13225.
- (8) Auffinger, P.; Hashem, Y. *Curr. Opin. Struct. Biol.* **2007**, *17*, 325.
- (9) Csaszar, K.; Spackova, N.; Stefl, R.; Sponer, J.; Leontis, N. B. *J. Mol. Biol.* **2001**, *313*, 1073.
- (10) Krasovska, M. V.; Sefcikova, J.; Spackova, N.; Sponer, J.; Walter, N. G. *J. Mol. Biol.* **2005**, *351*, 731.
- (11) McDowell, S. E.; Spackova, N.; Sponer, J.; Walter, N. G. *Biopolymers* **2007**, *85*, 169.
- (12) Ditzler, M. A.; Otyepka, M.; Sponer, J.; Walter, N. G. *Acc. Chem. Res.* **2010**, *43*, 40.
- (13) Denning, E. J.; Priyakumar, U. D.; Nilsson, L.; Mackerell, A. D. *J. Comput. Chem.* **2011**, *32*, 1929.
- (14) Yildirim, I.; Stern, H. A.; Kennedy, S. D.; Tubbs, J. D.; Turner, D. H. *J. Chem. Theory Comput.* **2010**, *6*, 1520.
- (15) Banas, P.; Hollas, D.; Zgarbova, M.; Jurecka, P.; Orozco, M.; Cheatham, T. E.; Sponer, J.; Otyepka, M. *J. Chem. Theory Comput.* **2010**, *6*, 3836.
- (16) Yildirim, I.; Stern, H. A.; Tubbs, J. D.; Kennedy, S. D.; Turner, D. H. *J. Phys. Chem. B* **2011**, *115*, 9261.
- (17) Chen, G.; Kierzek, R.; Yildirim, I.; Krugh, T. R.; Turner, D. H.; Kennedy, S. D. *J. Phys. Chem. B* **2007**, *111*, 6718.
- (18) Chen, X.; Kierzek, R.; Turner, D. H. *J. Am. Chem. Soc.* **2001**, *123*, 1267.
- (19) Petersheim, M.; Turner, D. H. *Biochemistry* **1983**, *22*, 256.
- (20) Petersheim, M.; Turner, D. H. *Biochemistry* **1983**, *22*, 264.
- (21) Petersheim, M.; Turner, D. H. *Biochemistry* **1983**, *22*, 269.
- (22) Kollman, P. A. *Chem. Rev.* **1993**, *93*, 2395.
- (23) 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.; Mathews, D. H.; Schafmeister, C.; Ross, W. S.; Kollman, P. A. *AMBER 9*; University of California San Francisco: San Francisco, CA, 2006.
- (24) Perez, A.; Marchan, I.; Svozil, D.; Sponer, J.; Cheatham, T. E.; Laughton, C. A.; Orozco, M. *Biophys. J.* **2007**, *92*, 3817.
- (25) Beaucage, S. L.; Caruthers, M. H. *Tetrahedron Lett.* **1981**, *22*, 1859.

- (26) Kierzek, R.; Caruthers, M. H.; Longfellow, C. E.; Swinton, D.; Turner, D. H.; Freier, S. M. *Biochemistry* **1986**, *25*, 7840.
- (27) Testa, S. M.; Disney, M. D.; Turner, D. H.; Kierzek, R. *Biochemistry* **1999**, *38*, 16655.
- (28) Xia, T.; SantaLucia, J., Jr.; Burkard, M. E.; Kierzek, R.; Schroeder, S. J.; Jiao, X.; Cox, C.; Turner, D. H. *Biochemistry* **1998**, *37*, 14719.
- (29) Varani, G.; Aboulela, F.; Allain, F. H. T. *Prog. Nucl. Magn. Reson. Spectrosc.* **1996**, *29*, 51.
- (30) Varani, G.; Tinoco, I. Q. *Rev. Biophys.* **1991**, *24*, 479.
- (31) Yildirim, I.; Stern, H. A.; Sponer, J.; Spackova, N.; Turner, D. H. *J. Chem. Theory Comput.* **2009**, *5*, 2088.
- (32) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision C.02; Gaussian, Inc.: Wallingford, CT, 2004.
- (33) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.
- (34) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327.
- (35) Lu, X. J.; Olson, W. K. *Nature Protoc.* **2008**, *3*, 1213.
- (36) Freier, S. M.; Burger, B. J.; Alkema, D.; Neilson, T.; Turner, D. H. *Biochemistry* **1983**, *22*, 6198.
- (37) Freier, S. M.; Sinclair, A.; Neilson, T.; Turner, D. H. *J. Mol. Biol.* **1985**, *185*, 645.
- (38) Richardson, J. S.; Schneider, B.; Murray, L. W.; Kapral, G. J.; Immormino, R. M.; Headd, J. J.; Richardson, D. C.; Ham, D.; Hershkovits, E.; Williams, L. D.; Keating, K. S.; Pyle, A. M.; Micallef, D.; Westbrook, J.; Berman, H. M. *RNA* **2008**, *14*, 465.
- (39) Schneider, B.; Moravek, Z.; Berman, H. M. *Nucleic Acids Res.* **2004**, *32*, 1666.
- (40) Baker, D. *Protein Sci.* **2010**, *19*, 1817.
- (41) Ponder, J. W.; Case, D. A. In *Protein Simulations*; Elsevier Academic Press: San Diego, Ca, **2003**; Vol. 66, p 27.
- (42) Salsbury, F. R. *Curr. Opin. Pharmacol.* **2010**, *10*, 738.
- (43) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y. B.; Wriggers, W. *Science* **2010**, *330*, 341.
- (44) Disney, M. D.; Childs, J. L.; Turner, D. H. *Biopolymers* **2004**, *73*, 151.
- (45) Childs-Disney, J. L.; Wu, M. L.; Pushechnikov, A.; Aminova, O.; Disney, M. D. *ACS Chem. Biol.* **2007**, *2*, 745.
- (46) Disney, M. D.; Labuda, L. P.; Paul, D. J.; Poplawski, S. G.; Pushechnikov, A.; Tran, T.; Velagapudi, S. P.; Wu, M.; Childs-Disney, J. L. *J. Am. Chem. Soc.* **2008**, *130*, 11185.
- (47) Pushechnikov, A.; Lee, M. M.; Childs-Disney, J. L.; Sobczak, K.; French, J. M.; Thornton, C. A.; Disney, M. D. *J. Am. Chem. Soc.* **2009**, *131*, 9767.
- (48) Zgarbova, M.; Otyepka, M.; Sponer, J.; Mladek, A.; Banas, P.; Cheatham, T. E.; Jurecka, P. *J. Chem. Theory Comput.* **2011**, *7*, 2886.

Cation– π and π – π Interactions in Aqueous Solution Studied Using Polarizable Potential Models

Esam A. Orabi[†] and Guillaume Lamoureux^{*}

Department of Chemistry and Biochemistry and Centre for Research in Molecular Modeling, Concordia University, Montréal, Québec H4B 1R6, Canada

ABSTRACT: Polarizable potential models for the interaction of Li^+ , Na^+ , K^+ , and NH_4^+ ions with benzene are parametrized based on ab initio quantum mechanical calculations. The models reproduce the ab initio complexation energies and potential energy surfaces of the cation– π dimers. They also reproduce the cooperative behavior of “stacked”, cation– π – π trimers and the anticooperative behavior of “sandwiched”, π –cation– π trimers. The NH_4^+ model is calibrated to reproduce the energy of the NH_4^+ – H_2O dimer and yields correct free energy of hydration and hydration structure without further adjustments. The models are used to investigate cation– π interactions in aqueous solution by calculating the potential of mean force between each of the four cations and a benzene molecule and by analyzing the organization of the solvent as a function of the cation–benzene separation. The results show that Li^+ and Na^+ ions are preferentially solvated by water and do not associate with benzene, while K^+ and NH_4^+ ions bind benzene with 1.2 and 1.4 kcal/mol affinities, respectively. Molecular dynamics simulations of NH_4^+ and of K^+ in presence of two benzene molecules in water show that cation– π and π – π affinities are mutually enhanced compared to the pairwise affinities, confirming that the cooperativity of cation– π and π – π interactions persists in aqueous solution.

1. INTRODUCTION

Cation– π interactions are noncovalent interactions between positively charged ions and the π electrons of aliphatic or aromatic compounds.^{1–5} Such pairings of cations and π systems have been the subject of multiple experimental^{6–14} and computational^{2–5,15–25} studies. Experimental studies have shown that cation– π interactions in the gas phase are competitive with some of the strongest noncovalent interactions.^{6,7,10–12} Analysis of high-resolution structures in the Protein Data Bank²⁶ shows that cation– π interactions are commonly found in proteins^{5,27} and at protein–protein²⁸ and protein–DNA interfaces.^{29,30} Cation– π interactions contribute to protein stability,^{31,32} protein–ligand interactions,^{1,33} and to molecular recognition in general.³⁴

Quantum mechanical (QM) calculations on cation– π complexes correlate strongly with experimental gas phase data.^{2–5,9–12,14–25} For instance, binding enthalpies of alkali metal ions with benzene calculated at state-of-the-art levels of theory have shown good agreement with experimental values.¹⁷ QM calculations—and molecular mechanics models that accurately reproduce QM-calculated properties—thus serve as a convenient tool for studying and understanding cation– π interactions. In particular, it has been shown that the dominant contributions to cation– π interactions are electrostatics and polarization: charge–quadrupole and charge–induced dipole, mainly.^{1,2,35} Other forces, such as dispersion and charge transfer, are much weaker. Electronic polarization is a determining factor, due to the strong electric field produced by the cation.^{15,16,18,22}

Cation– π interactions usually out-compete cation–water interactions in the gas phase. For example the enthalpy of formation of K^+ –benzene complex in gas phase is -19.2 kcal/mol, compared to -17.9 kcal/mol for K^+ – H_2O .³⁶ Cation– π

interactions are weaker in solution than in gas phase,¹⁹ due to the charge screening effect of the solvent and the high availability of water. Nevertheless, their existence in aqueous solution has been computationally^{37–39} and experimentally^{8,13} confirmed. While the computational and experimental literature on cation– π interactions in the gas phase is abundant, studies of the interactions in water are few, and further work is required for a detailed understanding of their stability in solution.

Computational studies of cation– π_2 complexes in which π systems are arranged in a stacked geometry^{21,25} show that cation– π and π – π interactions are cooperative: The presence of one interaction strengthens the other and results in a net increase of the complex stabilization energy.²⁵ It is however not clear how such cooperativity in gas phase translates in aqueous solution, where cation– π interactions are competing with cation–water interactions and where π – π interactions are stabilized by the hydrophobic effect. These are likely important considerations for the binding of cationic moieties to proteins.

Owing to the biological importance of cation– π interactions and the computational prohibition of QM calculations on these systems, computationally inexpensive yet accurate molecular models for these interactions are crucial. Since electronic polarization represents an important contribution to cation– π interactions, polarizable potential models are required.^{15,35}

In this work, we parametrize polarizable empirical force fields for the interaction of Li^+ , Na^+ , K^+ , and NH_4^+ with benzene as well as for the interaction of NH_4^+ with water. Electronic polarization in the systems is described using classical Drude oscillators.^{40–42} We apply these models to investigate cation–benzene

Received: August 15, 2011

Published: November 09, 2011

interactions in water and their interplay with benzene–benzene interactions. The $\text{NH}_4^+ - \text{H}_2\text{O}$ interaction model is validated by calculating the free energy of hydration and hydration structure of the ion.

We perform ab initio calculations (geometry optimizations and potential energy scans) on the four cation–benzene complexes as well as on the $\text{NH}_4^+ - \text{H}_2\text{O}$ complex. The calculated ab initio properties are then used to parametrize the polarizable potential models. Molecular dynamics (MD) simulations of M^+ –benzene (where M^+ is Li^+ , Na^+ , K^+ , or NH_4^+), $(\text{benzene})_2$, and $\text{NH}_4^+/\text{K}^+ - (\text{benzene})_2$ complexes in bulk water are performed using the polarizable models, in order to measure the strength of individual cation– π and $\pi - \pi$ interactions in water and to understand how one type of interaction affects the other.

2. METHODS

2.1. Ab Initio Calculations. The geometries of Li^+ , Na^+ , K^+ , and NH_4^+ in complex with water, benzene, and with the benzene dimer and trimer are optimized at the Møller–Plesset MP2/6-311++G(d,p) level with frozen core (FC) electrons, using Gaussian 09.⁴³ The interaction energies are corrected for basis set superposition error (BSSE) by the counterpoise method of Boys and Bernardi⁴⁴ (and referred to as E^{CP}). Similar calculations are performed on the water–benzene complex and for the benzene dimer and trimer. The optimization of Li^+ , Na^+ , K^+ , and NH_4^+ in complex with benzene and with water is performed without imposing any geometry constraints. The optimization of the complexes with benzene dimer and trimer is performed imposing C_{6v} symmetry for the alkali metal ions and C_{2v} symmetry for the NH_4^+ ion, such that the benzene molecules remain parallel and undisplaced. Optimization of benzene dimer and trimer is performed imposing D_{6h} symmetry as well. Although the cation– $(\text{benzene})_2$ and cation– $(\text{benzene})_3$ systems are not directly used to calibrate the polarizable models, they allow to further investigate the cooperativity between cation– π and $\pi - \pi$ interactions^{21,25} and to test the performance of the polarizable force fields in describing such cooperativity. Two conformations of the cation– $(\text{benzene})_2$ systems are studied: the cation–benzene–benzene “stack” conformation and the benzene–cation–benzene “sandwich” conformation. The sandwich structures are optimized without any constraints, in both the straight and bent conformations.

Potential energy surfaces (PESs) of the four cation–benzene dimers and of the $\text{NH}_4^+ - \text{H}_2\text{O}$ dimer are calculated at the MP2(FC)/6-311++G(d,p) level, and all interaction energies are corrected for BSSE. The surfaces are computed with the molecular fragments kept in their optimized gas phase geometries, calculated at the same level of theory. PESs for the alkali metal cations are calculated by scanning both the perpendicular and parallel displacement of the ion relative to the benzene plane. For ammonium ion in complex with benzene, two potential curves are calculated. The first curve is calculated by scanning the distance between the nitrogen atom of NH_4^+ (in its bidentate conformation) and the center of the benzene molecule (labeled X). The second curve is calculated by scanning the angle $\text{X} \cdots \text{N} - \text{H}$, which shows the interaction energy as a function of the orientation of the ion (unidentate, bidentate, or tridentate) on top of benzene surface. Two curves are calculated for the $\text{NH}_4^+ - \text{H}_2\text{O}$ complex by scanning the $\text{N} \cdots \text{O}$ distance in the unidentate conformation of the ion and the orientation of NH_4^+ relative

to O (unidentate, bidentate, or tridentate) at the optimal $\text{N} \cdots \text{O}$ distance.

2.2. Molecular Mechanical Calculations. **2.2.1. Potential Energy Function and Parametrization Strategy.** Molecular mechanics (MM) calculations are performed with the program CHARMM.⁴⁵ Polarizable models based on classical Drude oscillators^{40–42} are parametrized for the interaction of the four cations (Li^+ , Na^+ , K^+ , and NH_4^+) with benzene and for the interaction of NH_4^+ with water. For the interaction of Li^+ , Na^+ and K^+ with water, we use previously developed models.⁴⁶ Polarizability is introduced by attaching fictitious charged particles to all nonhydrogen atoms via a harmonic spring with force constant k_D . The partial charge of the polarizable atom, q , is distributed between the Drude particle and the atom core with the Drude particle charge q_D being determined from the atomic polarizability via the relation $\alpha = q_D^2/k_D$. The net charge of the atomic core is thus $q_c = q - q_D$. A separation d between the Drude particle and the polarizable atom results in an induced dipole moment $q_D d$. The electrostatic energy term in the additive potential energy function⁴⁷ is modified to include interactions between atomic cores and Drude particles. A term describing the self-energy of a polarizable atom [$1/2(k_D d^2)$] is also added to the potential energy function.⁴¹ The resulting potential energy function can be written as the following:^{47,48}

$$U(R) = U_{\text{ion}}(R) + U_{\text{ion-solvent}}(R) + U_{\text{solvent}}(R) \quad (1)$$

where

$$U_{\text{ion}}(R) = \frac{1}{2}k_D|\mathbf{r}_{\text{ion}} - \mathbf{r}_{D,\text{ion}}|^2 + \sum_{\text{NH bonds}} k_b(b - b_0)^2 + \sum_{\text{HNN angles}} k_\theta(\theta - \theta_0)^2 + \sum_{\text{HH pairs}} k_{\text{UB}}(s - s_0)^2 \quad (2)$$

in which the last three terms are for NH_4^+ only (b are NH bond lengths, θ are HNH angles, and s are HH distances), and where

$$U_{\text{ion-solvent}}(R) = \sum_{j=1}^N \sum_i \sum_s \left(\frac{q_{c,i}q_s}{|\mathbf{r}_i - \mathbf{r}_{js}|} + \frac{q_{D,i}q_s}{|\mathbf{r}_{D,i} - \mathbf{r}_{js}|} \right) + \sum_{j=1}^N \sum_i \sum_s E_{\text{min},is} \left[\left(\frac{R_{\text{min},is}}{|\mathbf{r}_i - \mathbf{r}_{js}|} \right)^{12} - 2 \left(\frac{R_{\text{min},is}}{|\mathbf{r}_i - \mathbf{r}_{js}|} \right)^6 \right] \quad (3)$$

in which N is the number of solvent molecules, i is the atomic site of the ion (Li, Na, K, N, H), and s is the solvent molecule site (atoms, lone pairs, Drude particles). $U_{\text{solvent}}(R)$, the third term in eq 1, is similarly obtained as the sum of bonded and nonbonded energy terms that correspond to interaction between atoms of the solvent molecules. Parameters in these equations and their definitions can be found in refs 47 and 48.

Parameters for the NH_4^+ potential function are obtained based on ab initio calculations on the ion and its complex with benzene. The NH_4^+ ion is modeled by five atomic sites and an auxiliary Drude particle attached to the nitrogen atom. Parameters b_0 , θ_0 , and s_0 are set according to the ab initio geometry of the monomer in gas phase. NH bond stretching energy is represented by harmonic terms, and although these bonds are kept rigid during the simulations, the force constant k_b is adjusted to reproduce ab initio frequencies of the stretching modes. Angle bending terms are adjusted to reproduce the distortion energy associated with the bending modes. Urey–Bradley (UB) energy terms⁴⁹ are added to improve vibration frequencies and to prevent large

distortions in the tetrahedral structure of the ion. The electrostatic parameters (atomic charges and polarizabilities) are determined from ab initio calculation. The atomic charges are fitted to reproduce the traceless quadrupole moment of the ion, and the polarizability of N is calculated from the trace of the polarizability tensor.

Lennard-Jones (LJ) parameters of N and H atoms of ammonium are optimized separately for the interaction with benzene and water. Parameters of benzene are taken from ref 49. Parameters for the alkali metal ions are taken from ref 46. An extra nonatomic site (X) at the center of the benzene ring is required in order to accurately model the interactions with Na^+ and NH_4^+ ions. This site mimics the electron density at the center of the benzene ring and the repulsive effect it has on the ions. It is electrically neutral and shows a LJ interaction only with Na^+ and with the N and H atoms of NH_4^+ . The X site was not required for lithium, likely due to the small size of the ion and the close contact it forms with the benzene ring. It was also not required for potassium, which is too large to “discriminate” the steric profile of the ring.

The general parametrization strategy of the polarizable force field based on Drude oscillators has been documented elsewhere.^{41,42,46,49} In the present work, the strength of the interaction of the four cations with benzene and of NH_4^+ with water is adjusted by optimizing the LJ parameters between specific pairs of atoms of the monomers. The NBFIX⁴⁵ option of CHARMM allows assigning pair-specific LJ parameters $E_{\text{min},is}$ and $R_{\text{min},is}$ that override the default values obtained from the Lorentz–Berthelot combination rules:

$$E_{\text{min},is} = \sqrt{E_{\text{min},i} E_{\text{min},s}} \quad \text{and} \quad R_{\text{min},is} = (R_{\text{min},i} + R_{\text{min},s})/2$$

Models for alkali metal–benzene interactions are optimized by adjusting pair-specific LJ parameters between the ions (Li^+ , Na^+ , and K^+) and the carbon atoms of benzene and between Na^+ and the X site of benzene. Model for NH_4^+ –benzene interaction is optimized by adjusting the LJ parameters of N and H atoms of NH_4^+ as well as their pair-specific parameters with the X site in benzene. The interaction of NH_4^+ with H_2O is optimized by adjusting pair-specific LJ parameters between N and H atoms of NH_4^+ and oxygen atom of H_2O .

Parameter optimization uses the ab initio properties (complexation energies, geometries, and PESs) as targets and is performed in two steps. The first optimization step is to reproduce the ab initio PESs around their minimum. In this step the coordinates of the complex are kept rigid (at the values used for the ab initio potential energy scans), and the LJ parameters are adjusted to minimize the following error function:

$$\chi^2 = \sum_k (E_k^{\text{CP}} - E_k^{\text{MM}})^2 e^{-E_k^{\text{CP}}/k_{\text{B}}T}$$

where E^{CP} is the BSSE-corrected ab initio interaction energy, E^{MM} is the interaction energy from the polarizable model, k_{B} is Boltzmann constant, and T is the standard temperature (298.15 K). Index k represents the grid points on the potential energy surface. Minimization of this function leads to the best overall agreement between the interaction energies calculated from the Drude polarizable model (E^{MM}) and the corresponding ab initio values (E^{CP}). The sum of squared errors is weighted by a Boltzmann factor, which has the effect of increasing the importance of the low-energy conformations and ensures that the bottom of the energy surface is well reproduced.

The parameters obtained from this procedure are subjected to a second optimization step, in which the geometry and the

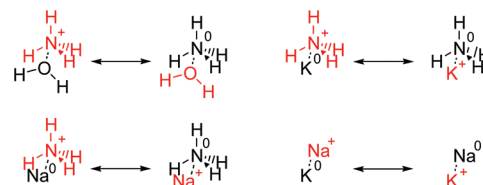


Figure 1. Solute transformations involved in the free energy calculations. Fragments in red are “real” while those in black are “dummy”. The dashed line represents the harmonic link.

interaction energy of the complex calculated without imposing geometry constraints (except fixing the bonds to H atoms using the SHAKE algorithm)⁵⁰ are fitted to the corresponding ab initio results. While the purpose of the first optimization step is to find the set of LJ parameters that gives the best global accuracy, the refined parameters resulting from the second step are more reliable, as they describe the geometry and the energetics of the complex under simulation conditions.

2.2.2. MD Simulations. MD simulations are performed in order to investigate cation– π , π – π , and cation– π_2 interactions in water as well as the hydration structure of NH_4^+ . All simulations are performed in the NPT ensemble at $T = 298.15$ K and $p = 1$ atm, with cubic periodic boundary conditions. Single ions are solvated in 250 water molecules; ion–benzene and benzene–benzene pairs and ion–benzene–benzene triples are solvated in 600 water molecules. The SWM4-NDP polarizable water model⁵¹ is used for all simulations with a mass of 0.4 au on the auxiliary Drude particles and a force constant $k_{\text{D}} = 1000$ kcal/mol/Å² for the atom–Drude coupling. Electrostatic interactions are computed using the particle-mesh Ewald method,⁵² with $\kappa = 0.34$ for the charge screening and a 1.0 Å grid spacing with fourth-order splines for the mesh interpolation. The real-space interactions (LJ and electrostatic) are cut off at 15 Å, and the long-range contribution from the LJ term is introduced as an average density-dependent term.⁵³ The temperature of the system is controlled with a two-thermostat algorithm, where atoms are kept at room temperature (298.15 K) and auxiliary Drude particles are kept at low temperature (1 K) to ensure self-consistent dipole induction.⁴¹ The equations of motion are integrated using a 1 fs time step, with all bonds involving hydrogen atoms kept at their reference lengths using the SHAKE algorithm.⁵⁰

2.2.3. Free Energy Calculations. The polarizable model for the NH_4^+ – H_2O complex is validated by calculating the free energy of hydration of an NH_4^+ ion relative to H_2O , Na^+ , and K^+ . Free energy calculations are performed following the thermodynamic integration (TI) simulation protocol established previously.⁴⁸ In particular, the relative hydration free energy ($\Delta\Delta G_{\text{hydr}}$) of solutes A and B is evaluated from the conventional thermodynamic cycle for solute transformation in water:

$$\Delta\Delta G_{\text{hydr}}(A \rightarrow B) \equiv \Delta G_{\text{hydr}}(B) - \Delta G_{\text{hydr}}(A) = \Delta G_{\text{mut}}^{\text{wat}}$$

where $\Delta G_{\text{mut}}^{\text{wat}}$ is the relative free energy for the alchemical $A \rightarrow B$ “mutation” in water.

To maintain a constant number of interaction sites throughout the transformation, special hybrid residues are used (see Figure 1) in which A and B solutes are linked through their heavy atoms via a weak harmonic bond of force constant 5 kcal/mol/Å². These residues are formed by tethering one original “real” ion with a “dummy” ion having no interactions with the real particles. The mutation simply involves “turning off” the nonbonded

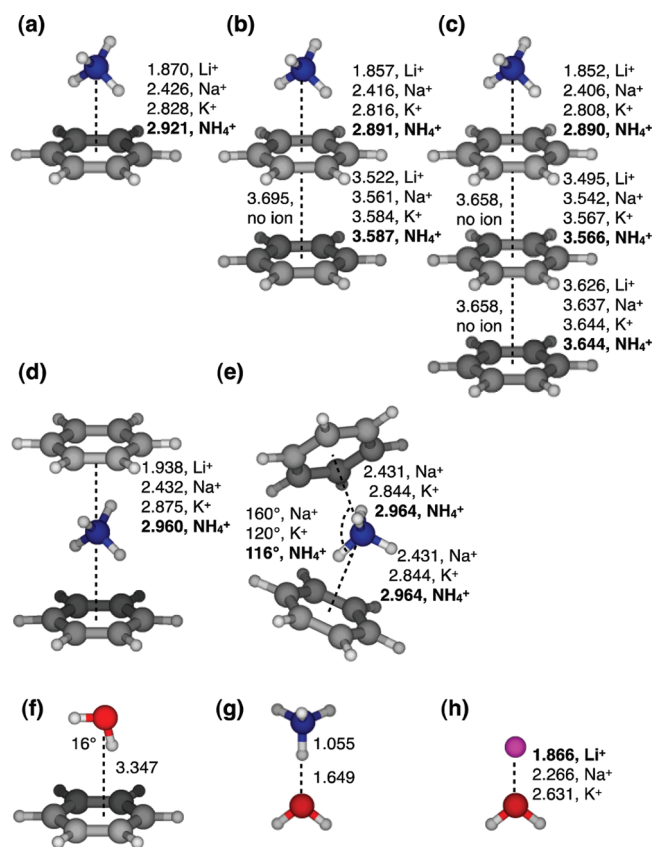


Figure 2. Optimized geometries at MP2(FC)/6-311++G(d,p) level of theory for the (a) cation–benzene, (b) cation–(benzene)₂, (c) cation–(benzene)₃, (d) benzene–cation–benzene, (e) benzene–cation–benzene bent sandwich (unstable for Li⁺), (f) water–benzene, (g) ammonium–water, and (h) alkali ion–water complexes. The structures of panels (a–e) are illustrated with NH₄⁺ complexes, but corresponding parameters for the lithium, sodium, and potassium complexes are reported.

parameters of the real fragment while “turning on” those of the dummy fragment. The ligand transformation is performed in 17 steps, controlled by a scaling parameter λ which takes the following values: 0, 0.002, 0.005, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.995, 0.998, and 1. Each λ window is equilibrated for 150 ps followed by subsequent data collection for 350 ps.

2.2.4. Potential of Mean Force Calculations. Potentials of mean force (PMFs) between each cation (Li⁺, Na⁺, K⁺, or NH₄⁺) and one benzene molecule and between two benzene molecules are calculated using umbrella sampling. The distance between the centers of mass (CMs) of the reactants is used as a reaction coordinate, and a harmonic potential of force constant 10 kcal/mol/Å² is applied to bias the sampling. The reaction coordinate is sampled using 0.5 Å separated windows, and each window is simulated for 2 ns. The unbiased PMF is reconstructed using the weighted histogram analysis method (WHAM),^{54,55} and the radial variation in the entropy of the solute pairs is taken into account by adding a $2k_B T \ln(R)$ correction term to the PMF.⁵⁶

3. RESULTS AND DISCUSSION

3.1. Ab Initio Interaction Energies. The optimized geometries of all studied complexes and some of their structural parameters are presented in Figure 2. BSSE-corrected and -uncorrected complexation energies (E^{CP} and E , respectively) and equilibrium

Table 1. Ab Initio Complexation Energies and Equilibrium Distances (R_1 , R_2 , R_3) Calculated at the MP2(FC)/6-311++G(d,p) Level of Theory and Corresponding Energies Calculated Using the Polarizable Models (E^{MM})^a

complex	E	E^{CP}	R_1	R_2	R_3	E^{MM}
benz–Li ⁺	–38.84	–34.89	1.870	–	–	–35.14
benz–Na ⁺	–24.00	–21.08	2.426	–	–	–21.04
benz–K ⁺	–19.58	–17.14	2.828	–	–	–17.01
benz–NH ₄ ⁺	–19.78	–17.58	2.921	–	–	–17.56
(benz) ₂ –Li ⁺	–49.00	–41.48	1.857	3.522	–	–40.59
(benz) ₂ –Na ⁺	–32.92	–26.65	2.416	3.561	–	–25.66
(benz) ₂ –K ⁺	–28.31	–22.23	2.816	3.584	–	–21.25
(benz) ₂ –NH ₄ ⁺	–28.67	–22.82	2.891	3.587	–	–21.84
benz–Li ⁺ –benz	–71.88	–60.57	1.938	–	–	–62.47
benz–Na ⁺ –benz	–46.50	–38.86	2.432	–	–	–40.16
	–46.47 ^c	–38.85 ^c	2.431 ^c	–	–	–39.04
benz–K ⁺ –benz	–37.56	–32.09	2.875	–	–	–33.09
	–38.47 ^c	–32.26 ^c	2.844 ^c	–	–	–33.07
benz–NH ₄ ⁺ –benz	–37.33	–32.17	2.960	–	–	–32.70
	–38.58 ^c	–31.80 ^c	2.964 ^c	–	–	–31.59
(benz) ₃ –Li ⁺	–56.12	–44.72	1.852	3.495	3.626	–43.62
(benz) ₃ –Na ⁺	–39.73	–29.63	2.406	3.542	3.637	–28.50
(benz) ₃ –K ⁺	–34.98	–25.10	2.808	3.567	3.644	–24.00
(benz) ₃ –NH ₄ ⁺	–35.37	–25.69	2.890	3.566	3.644	–24.63
(benz) ₂	–4.94	–1.79	–	3.695	–	–2.07
(benz) ₃	–10.67	–3.62	–	3.658	3.658	–4.20
benz–H ₂ O	–4.52	–2.43	3.347	–	–	–2.68 ^b
H ₂ O–Li ⁺	–35.50	–33.40	1.866	–	–	–35.92 ^b
H ₂ O–Na ⁺	–24.67	–23.09	2.266	–	–	–24.64 ^b
H ₂ O–K ⁺	–18.93	–17.88	2.631	–	–	–17.90 ^b
H ₂ O–NH ₄ ⁺	–22.16	–20.27	2.704	–	–	–20.28

^a E : uncorrected; and E^{CP} : BSSE-corrected. All energies in kcal/mol and all distances in Å. ^b Calculated using the original polarizable models.^{46,49,51} ^c Bent sandwich geometry of the complex (see Figure 2e). The Li⁺ complex is unstable.

distances (R_1 , R_2 , and R_3) are reported in Table 1. The complexation energies and equilibrium distances for the (benzene)₂, (benzene)₃, and benzene–H₂O complexes are also reported in Table 1. The three equilibrium distances, R_1 , R_2 , and R_3 , represent CM separations between the cation and the closest benzene molecule, between the closest and second closest benzene molecules, and between the second and third closest benzene molecules, respectively. R_1 is also assigned to the CM separation between the cations and water and between water and benzene. As reported in Table 1, the MP2(FC)/6-311++G(d,p) interactions energies of Li⁺, Na⁺, K⁺, and NH₄⁺ with benzene monomer are –34.89, –21.08, –17.14, and –17.58 kcal/mol, respectively. Although these values are simple interaction energies (neglecting thermodynamic contributions), they are comparable to the corresponding experimental gas phase binding enthalpies (at 298 K) of –39.3 ± 3.2, –22.5 ± 1.5, –17.7 ± 1.0, and –19.3 ± 1.0 kcal/mol.^{7,10}

For alkali cation complexes, the interaction energy decreases (that is, becomes less negative) while R_1 increases on going from Li⁺ to K⁺ (see Table 1), which can be attributed to the increase of the cation size. For a given cation, the interaction energy increases on going from benzene monomer to trimer complexes, while R_1 , R_2 , and R_3 decrease (see Table 1 and Figure 2a, b, and c).

Table 2. BSSE-Corrected Complexation Energies Calculated at the MP2(FC)/6-311++G(d,p) Level and Corresponding Interaction Energies Calculated Using the Polarizable Models (in parentheses)^a

complex	E_{tot}	$E_{\text{M-B}_1}$	$E_{\text{M-B}_2}$	$E_{\text{M-B}_3}$	$E_{\text{B}_1-\text{B}_2}$	$E_{\text{B}_2-\text{B}_3}$	$E_{\text{B}_1-\text{B}_3}$	E_{coop}
benz-Li ⁺	-34.89 (-35.14)	-34.89 (-35.14)	-	-	-	-	-	-
benz-Na ⁺	-21.08 (-21.04)	-21.08 (-21.04)	-	-	-	-	-	-
benz-K ⁺	-17.14 (-17.01)	-17.14 (-17.01)	-	-	-	-	-	-
benz-NH ₄ ⁺	-17.58 (-17.56)	-17.58 (-17.56)	-	-	-	-	-	-
(benz) ₂ -Li ⁺	-41.48 (-40.59)	-34.82 (-34.93)	-4.26 (-2.90)	-	-1.10 (-0.63)	-	-	-1.30 (-2.13)
(benz) ₂ -Na ⁺	-26.65 (-25.66)	-21.02 (-20.97)	-3.12 (-2.15)	-	-1.35 (-1.00)	-	-	-1.16 (-1.54)
(benz) ₂ -K ⁺	-22.23 (-21.25)	-17.09 (-17.02)	-2.59 (-1.84)	-	-1.47 (-1.18)	-	-	-1.08 (-1.21)
(benz) ₂ -NH ₄ ⁺	-22.82 (-21.84)	-17.56 (-18.24)	-2.57 (-1.81)	-	-1.48 (-1.20)	-	-	-1.21 (-0.59)
benz-Li ⁺ -benz	-60.57 (-62.47)	-34.79 (-34.66)	-34.79 (-34.66)	-	-1.90 (-2.05)	-	-	10.91 (8.90)
benz-Na ⁺ -benz	-38.86 (-40.16) -38.85 ^b (-39.04) ^b	-20.99 (-20.98) -20.99 ^b (-20.61) ^b	-20.99 (-20.98) -20.99 ^b (-20.61) ^b	-	-0.75 (-1.04) -0.89 ^b (-1.19) ^b	-	-	3.87 (2.82) 4.02 ^b (3.37) ^b
benz-K ⁺ -benz	-32.09 (-33.09) -32.26 ^b (-33.07) ^b	-17.17 (-17.12) -17.15 ^b (-17.10) ^b	-17.17 (-17.12) -17.15 ^b (-17.10) ^b	-	-0.20 (-0.44) -0.83 ^b (-1.14) ^b	-	-	2.45 (1.59) 2.87 ^b (2.27) ^b
benz-NH ₄ ⁺ -benz	-32.17 (-32.07) -31.80 ^b (-31.59) ^b	-17.39 (-17.37) -16.99 ^b (-16.23) ^b	-17.42 (-17.35) -16.99 ^b (-16.23) ^b	-	-0.13 (-0.35) -0.78 ^b (-1.14) ^b	-	-	2.77 (3.00) 2.96 ^b (2.01) ^b
(benz) ₃ -Li ⁺	-44.72 (-43.62)	-34.80 (-34.92)	-4.35 (-2.97)	-0.93 (-0.60)	-0.93 (-0.36)	-1.63 (-1.45)	-0.001 (-0.12)	-2.08 (-3.20)
(benz) ₃ -Na ⁺	-29.63 (-28.50)	-20.99 (-20.91)	-3.18 (-2.20)	-0.76 (-0.49)	-1.26 (-0.85)	-1.66 (-1.50)	-0.002 (-0.12)	-1.78 (-2.43)
(benz) ₃ -K ⁺	-25.10 (-24.00)	-17.09 (-17.00)	-2.64 (-1.87)	-0.66 (-0.44)	-1.39 (-1.06)	-1.68 (-1.53)	-0.003 (-0.11)	-1.64 (-1.99)
(benz) ₃ -NH ₄ ⁺	-25.69 (-24.63)	-17.55 (-18.23)	-2.60 (-1.84)	-0.65 (-0.43)	-1.39 (-1.05)	-1.68 (-1.54)	-0.003 (-0.12)	-1.82 (-1.42)

^a E_{tot} is the total complexation energy, $E_{\text{A-B}}$ are complexation energies of the different fragment pairs, and E_{coop} is the cooperativity [see eq 4]. All energies in kcal/mol. ^b Bent sandwich geometry of the complex (see Figure 2e). The Li⁺ complex is unstable.

The values of R_1 reported in Table 1 for Li⁺, Na⁺, and K⁺ complexes with benzene (1.870, 2.426, and 2.828 Å, respectively) are in close agreement with the distances calculated by Feller et al.¹⁷ at the MP2/CBS level of theory (1.879, 2.390, and 2.786 Å, respectively).

Table 1 shows that cations are always closer to the benzene molecule (R_1 shorter) in cation-dimer complexes than in cation-monomer complexes and that two benzene molecules are always closer (R_2 shorter) in cation-dimer complexes than in the benzene-dimer. This indicates that cation- π and π - π interactions stabilize one another and that cooperativity between the two interactions contributes to the overall stabilization of the system.²¹

By comparison, the sandwich benzene-cation-benzene complexes (see Figure 2d and e), while more stable than

the cation-benzene-benzene conformers, display *competitive* cation- π interactions. For instance, the straight sandwich ammonium complex has a total complexation energy less than twice that of the NH₄⁺-benzene pair (-32.17 kcal/mol, compared to $2 \times -17.58 = -35.16$ kcal/mol), and the cation-benzene distances are larger (2.960 Å, compared to 2.921 Å for the NH₄⁺-benzene pair). The bent sandwich conformation, almost iso-energetic to the straight conformation, displays similar competitiveness. It is important for a molecular model to reproduce these effects, as the cooperation between cation- π and π - π interactions and the competition between two cation- π interactions are not expected to play out the same way in aqueous solution as in gas phase.

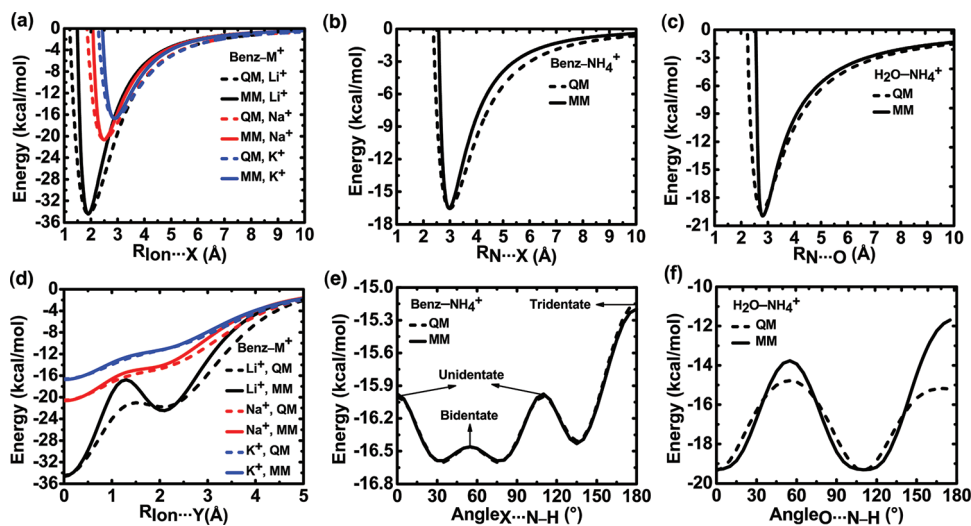


Figure 3. Potential energy curves for benzene– M^+ ($M^+ = \text{Li}^+, \text{Na}^+, \text{K}^+, \text{or } \text{NH}_4^+$) and $\text{H}_2\text{O}-\text{NH}_4^+$ complexes from ab initio MP2(FC)/6-311++G (d,p) calculations (dashed line) and from polarizable model (solid line). The following coordinates are scanned: (a) in benzene– M^+ complex, $X \cdots M^+$ distance in the direction perpendicular to the benzene plane (X is the center of the benzene ring); (b) in benzene– NH_4^+ complex, $X \cdots \text{N}$ distance between the benzene center and the nitrogen atom of NH_4^+ in its bidentate conformation; (c) in $\text{H}_2\text{O}-\text{NH}_4^+$ complex, $\text{O} \cdots \text{N}$ distance; (d) in benzene– M^+ complex, $Y \cdots M^+$ distance in the direction parallel to the benzene plane (Y is the equilibrium position of the ion; see Figure 2a), going toward the C–C bond center; (e) in benzene– NH_4^+ complex, $X \cdots \text{N}-\text{H}$ angle for NH_4^+ on top of benzene at $X \cdots \text{N}$ distance = 3.0 Å; and (f) in $\text{H}_2\text{O}-\text{NH}_4^+$ complex, $\text{O} \cdots \text{N}-\text{H}$ angle at $\text{O} \cdots \text{N}$ distance = 2.7 Å.

We evaluate E_{coop} , the cooperation energy of a complex, as the difference between E_{tot} , the total complexation energy, and the sum of all pairwise interaction energies in the complex:

$$E_{\text{coop}} = E_{\text{tot}} - \left(\sum_m E_{M-B_m} + \sum_{m < m'} E_{B_m-B_{m'}} \right) \quad (4)$$

where m and m' label the benzene molecules (see Table 2). E_{M-B_m} is the complexation energy of the different ion–benzene pairs and $E_{B_m-B_{m'}}$ is the complexation energy of the different benzene pairs—whether they are in contact or not. These energies are calculated at the geometry obtained from the optimization of the whole complex and corrected for BSSE.

Table 2 shows that the sign and the magnitude of E_{coop} depends on the nature of the complex. Parallel stacked cation–(benzene)₂ and cation–(benzene)₃ complexes show negative E_{coop} , which indicates that the two interactions strengthens one another.²⁵ The positive E_{coop} observed in the sandwich complexes of the benzene dimer indicates on the other hand that the two interactions are competitive.

This cooperative or anticooperative behavior is related to the polarization of the benzene molecules. Table 2 also reports the interaction energies calculated using the optimized Drude models (see Section 3.3). In the stacked conformation of the NH_4^+ –(benzene)₂ complex, the polarizable model induces a dipole of +2.53 D in the first benzene ring and of +0.60 D in the second. Those two dipoles are parallel and result in a stabilization of the complex by 0.59 kcal/mol (compared to 1.21 kcal/mol from the ab initio calculations). In the straight sandwich conformation, the ion induces antiparallel dipoles of +2.01 and –2.01 D, which destabilizes the complex by 3.00 kcal/mol (compared to 2.77 kcal/mol from the ab initio calculations). This behavior cannot be reproduced with a conventional, nonpolarizable force field.

3.2. Ab Initio Potential Energy Surfaces. Ab initio potential energy curves for Li^+ , Na^+ , K^+ , and NH_4^+ in complex with benzene monomer and for NH_4^+ in complex with H_2O are reported in Figure 3, along with the corresponding curves

obtained from the optimized Drude models (see Section 3.3). Two curves are calculated for the interactions of the alkali cations with the benzene monomer (Figure 3a and d). Curve 3a is calculated by positioning the cation on top of the benzene center, along the six-fold symmetry axis, and by scanning the distance R between the cation and the ring centroid (site X) from 1.0 to 10.0 Å. This curve indicates that the depth and the extent of the potential energy well depends on the size of the cation and on its ability to approach the electron cloud of benzene.²⁰ Curve 3d is calculated by positioning the alkali cations on top of the benzene center at the equilibrium separation distances (called site Y) and scanning the movement of the cations parallel to the benzene ring, going toward the C–C bond center. This curve confirms that, although the interaction energy decreases as the cation moves away from the benzene center, the complex remains stable.²⁴

Two potential curves are calculated for ammonium–benzene complex (Figure 3b and e). Curve 3b is calculated by scanning the $X \cdots \text{N}$ distance with ammonium in the bidentate orientation. Curve 3e is generated by scanning the $X \cdots \text{N}-\text{H}$ angle and describes the interaction energy of the complex as a function of ammonium orientation: unidentate (0° and 109°), bidentate (55°), and tridentate (180°). This curve shows that the stability of the different ammonium conformers follows the order bidentate > unidentate > tridentate. The global minimum conformer however displays an angle $X \cdots \text{N}-\text{H}$ of 35° or 75° , corresponding to an ammonium orientation between the exact unidentate and bidentate conformations.

Two potential curves are calculated for the NH_4^+ – H_2O complex (Figure 3c and f). Curve 3c shows the scan of the $\text{N} \cdots \text{O}$ distance in the ammonium unidentate orientation, from 2.0 to 10.0 Å. Curve 3f shows the scan of angle $\text{O} \cdots \text{N}-\text{H}$ from 0° to 180° , so as to investigate the relative stability of the unidentate (0° and 109°), bidentate (55°), and tridentate conformers (180°). According to the QM calculations, the

Table 3. Pair-Specific LJ Parameters for the Interactions of Li^+ , Na^+ , K^+ , and NH_4^+ with Benzene and Water^a

ion	<i>i</i>	E_{min} (kcal/mol)	$R_{\text{min}}/2$ (Å)	ion–benzene interaction				ion–water interaction	
				$E_{\text{min},iC}$ (kcal/mol)	$R_{\text{min},iC}$ (Å)	$E_{\text{min},iX}$ (kcal/mol)	$R_{\text{min},iX}$ (Å)	$E_{\text{min},io}$ (kcal/mol)	$R_{\text{min},io}$ (Å)
NH_4^+	N	2.387030	1.306271	0.4058387 ^b	3.3962713 ^b	0.1470587	3.5005950	0.1018465	3.7592014
	H	0.003998	1.087051	0.0109515 ^b	2.3562713 ^b	0.0060183	3.2808392	0.0092367	2.8848120
Li^+	Li	0.030000	1.100000	0.0644005	3.1950579	0.0	0.0	0.0795506 ^c	2.8869290 ^c
Na^+	Na	0.0315100	1.461680	0.2004369	3.3592376	0.0099919	3.6398984	0.0815280 ^c	3.2486090 ^c
K^+	K	0.1419265	1.686652	0.4266716	3.5744944	0.0	0.0	0.1730273 ^c	3.4735811 ^c

^a Parameters with hydrogen atoms of benzene and water are obtained using the Lorentz–Berthelot mixing rules. ^b Obtained from benzene parameters⁴⁹ using the Lorentz–Berthelot mixing rules. ^c Obtained from alkali cations⁴⁰ and water⁵¹ parameters using the Lorentz–Berthelot mixing rules.

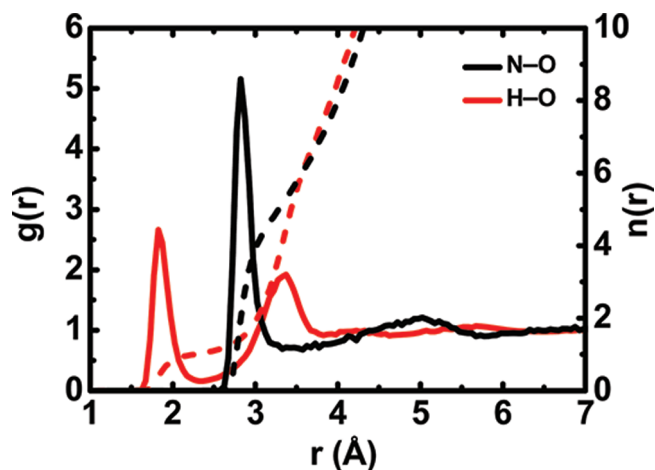


Figure 4. N–O and H–O radial distribution functions (solid lines, scale on left) and running integration numbers (dashed lines, scale on right) of NH_4^+ in water at 298.15 K.

stability of the conformers follows the order unidentate > bidentate ~ tridentate.

3.3. Optimized Force Field. Equilibrium structural parameters for the potential model of NH_4^+ (see eq 2) are obtained from ab initio optimization of the ion in the gas phase at the MP2/6-311++G(d,p) level of theory: $b_0 = 1.024$ Å, $\theta_0 = 109.47^\circ$, and $s_0 = 1.673$ Å. Bond, angle, and UB force constants are fitted in CHARMM⁴⁵ based on ab initio calculated IR frequencies of gaseous NH_4^+ ($\nu = 3 \times 1496, 2 \times 1734, 3413$, and 3×3547 cm^{-1}). Parameters $k_b = 470$ kcal/mol/Å², $k_\theta = 25$ kcal/mol/rad², and $k_{\text{UB}} = 9$ kcal/mol/Å² are chosen because they yield comparable IR frequencies ($\nu = 3 \times 1716, 2 \times 1940, 3461$, and 3×3546 cm^{-1}) and maintain structural stability of the ion during MD simulations. Although these frequencies are overestimating the ab initio bending vibrational frequencies of the ion (1496 and 1734 cm^{-1}), they reproduce the ab initio angle bending energies with less than 20% error. Since ab initio calculations of the hydration of ammonium in water clusters show that the HNH angles do not systematically bend by more than a fraction of a degree,⁵⁷ this represents an error of less than 0.01 kcal/mol.

Nonbonded parameters (atomic charges, polarizability, and LJ parameters) of the polarizable ammonium model are determined based on ab initio calculations on the gaseous ion and its complex with benzene and found to be $q(\text{H}) = 0.64413$ e, $q(\text{N}) = -1.57652$ e, $\alpha_{\text{N}} = 1.1966$ Å³, $E_{\text{min}}(\text{N}) = 2.387030$ kcal/mol, $R_{\text{min}}(\text{N})/2 = 1.306271$ Å, $E_{\text{min}}(\text{H}) = 0.003998$ kcal/mol, and $R_{\text{min}}(\text{H})/2 = 1.087051$ Å.

Table 4. Relative Hydration Free Energies (in kcal/mol) As Calculated from TI/MD Simulations in Bulk Water and Corresponding Experimental Values

mutation	$\Delta G_{\text{mut}}^{\text{wat}}$	experiment
$\text{NH}_4^+ \rightarrow \text{H}_2\text{O}$	61.7	61.8 ^{62a} , 65.6 ^{60a}
$\text{NH}_4^+ \rightarrow \text{Na}^+$	-18.6	-18.1, ⁵⁹ -19.1 ⁶²
$\text{NH}_4^+ \rightarrow \text{K}^+$	-1.2	-0.5, ⁵⁹ -2.4 ⁶²
$\text{Na}^+ \rightarrow \text{K}^+$	16.7	16.7, ⁶² 17.2, ⁶³ 17.6 ^{59,61}

^a Calculated using -6.32 kcal/mol as the experimental hydration free energy of water.⁶⁴

Optimized pair-specific LJ parameters for the interaction of the four cations (Li^+ , Na^+ , K^+ , and NH_4^+) with benzene and for $\text{NH}_4^+ - \text{H}_2\text{O}$ interaction are listed in Table 3. These parameters are first optimized based on the ab initio PESs and then refined to reproduce the ab initio geometry and interaction energy in the global minimum complex (see columns “ E^{CP} ” and “ E^{MM} ” of Table 1). The models also reproduce the ab initio PESs, as shown in Figure 3. Because of the intrinsic limitations of LJ potentials at reproducing both the position of the energy minimum and its curvature, the molecular models that yield correct binding energy, and equilibrium distance are systematically underestimating long-range interactions (see Figure 3a and b). For this reason, the interaction of the four cations with the benzene dimer in the stacked conformation (for which the second benzene molecule is about 6 Å away from the ion) is underestimated by about 1 kcal/mol. However, this systematic error does not increase with the stacking of a third benzene molecule, as the deviation between E^{MM} and E^{CP} becomes negligible at larger distances (see Figure 3a and b).

3.4. Hydration of NH_4^+ . The optimized model for $\text{NH}_4^+ - \text{H}_2\text{O}$ interaction reproduces the ab initio calculated complexation energy and PESs (see Table 1 and Figure 3c and f). To further validate the model, the solvation structure of the ion in water and its free energy of hydration relative to H_2O , Na^+ , and K^+ are calculated.

The solvation structure of the ammonium ion is investigated from the analysis of the last 7 ns of a 10 ns MD simulation of one ion solvated in 250 SWM4-NDP water molecules. The pair correlation functions $g_{\text{NO}}(r)$ and $g_{\text{HO}}(r)$ (where N and H refer to NH_4^+) are reported in Figure 4. Function $g_{\text{NO}}(r)$ shows a maximum at 2.85 Å and a minimum at 3.37 Å. Integration up to this minimum yields a coordination number of 5.3, in excellent agreement with ab initio MD studies,⁵⁸ which report a coordination number of 5.3 as well. Function $g_{\text{HO}}(r)$ shows a first peak at 1.85 Å and a minimum at 2.36 Å. Integration yields a coordination number of 1.05, corresponding to one water molecule

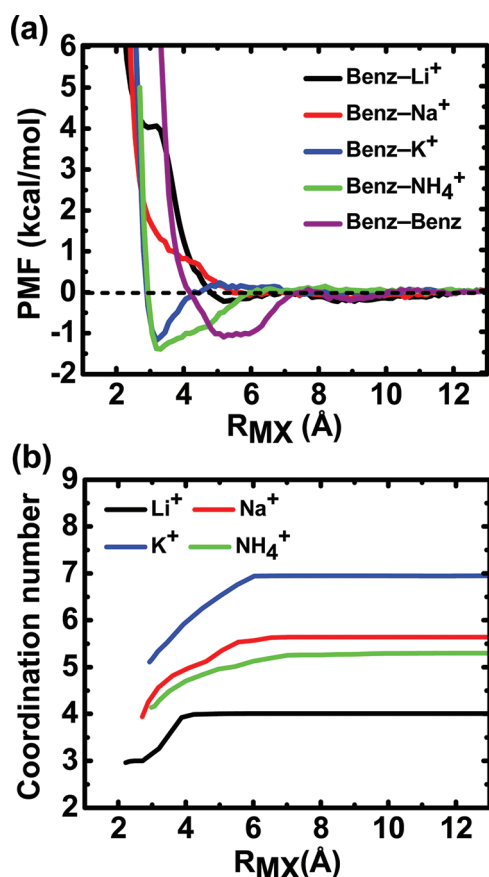


Figure 5. (a) PMFs between the centers of ions and benzene and between the centers of two benzene molecules in water. (b) Water-coordination number of Li⁺, Na⁺, K⁺, and NH₄⁺ ions as a function of their center-of-mass separation from a benzene molecule.

hydrogen-bonded to every proton of NH₄⁺, forming a tetrahedral structure around the ion. This suggests that the additional 1.3 water molecules found inside the N–O coordination sphere are much more mobile, in agreement with previous results from ab initio MD simulations.⁵⁸

Reliable simulations of ions in aqueous solution require models that reproduce their free energy of hydration (ΔG_{hyd}). To further validate the NH₄⁺–H₂O potential model, we calculate the change in free energy of hydration associated with mutation of NH₄⁺ to H₂O, Na⁺, or K⁺. As a control for the “hybrid residue” method used (see Section 2.2.3), the change in free energy for mutating Na⁺ to K⁺ is also calculated and compared to published values obtained using a different protocol.⁴⁶ The results are reported in Table 4, along with corresponding experimental and computational data.^{59–64} On the basis of multiple runs (forward and backward), the error on the calculated values is of the order of 0.1 kcal/mol.

The data in Table 4 show good agreement between the calculated and experimental data. Taking into account the hydration free energy of the SWM4-NDP water molecule, -5.9 ± 0.1 kcal/mol,⁵¹ this results in a hydration free energy of NH₄⁺ equal to -67.6 kcal/mol, in agreement with the experimental value of -68.1 kcal/mol reported by Marcus.⁶² It should be noted that the NH₄⁺–H₂O interaction model was not optimized to reproduce these experimental results; parametrization was aimed at reproducing the ab initio properties of the NH₄⁺–H₂O complex only. This further confirms the transferability of the parametrized NH₄⁺–H₂O model

from gas phase to aqueous phase. More importantly, the model reproduces the hydration free energy relative to Na⁺ and K⁺ (see Table 4), which ensures that the affinities for both water and benzene are correctly represented across the ion series.

3.5. Cation– π Interactions in Aqueous Solution. Compared to the gas phase interaction energies reported in Table 1, cation– π interactions are much weaker in aqueous solution,^{19,37–39} due to competing interactions with water. In the gas phase, the interaction energies of the alkali cations with benzene follow the order Li⁺ > Na⁺ > K⁺ (see Table 1). In water, the cation– π affinity is reported^{1,65} to show the reverse order: K⁺ \gg Na⁺ \sim Li⁺.

The binding affinity of the four studied cations with benzene in water is estimated from PMF calculations (see Figure 5a). Our finding for the binding free energies of Li⁺, Na⁺, and K⁺ with benzene in water is in agreement with the expected trend.⁶⁵ Li⁺ and Na⁺ do not associate with benzene in presence of water, as evidenced by the absence of a free energy minimum near the gas phase calculated equilibrium distances. For Li⁺, the shallow minimum (-0.2 kcal/mol) observed at CM separation of 5.1 Å can be interpreted as a weak interaction of benzene with the “dressed”, tetraqua Li⁺ ion. At that distance, the interaction energy of Li⁺ with benzene, as calculated from the PES, is -4.7 kcal/mol (see Figure 3a). Benzene in the second solvation shell of the ion will thus be stabilized by the interaction with ion-coordinated water molecules in addition to the long-distance interaction with the cation. K⁺ and NH₄⁺, on the other hand, bind benzene in water with energies of -1.2 and -1.4 kcal/mol at equilibrium CM separations of 3.2 and 3.3 Å, respectively. These equilibrium separations are 0.4 Å longer than the gas phase-calculated distances (see Table 1) but are nevertheless consistent with a direct coordination of the ions with the benzene molecule.

A benzene molecule coming in direct contact with a hydrated ion results in the expulsion of a number of water molecules from the first solvation shell. For this reason small, strongly solvated ions, such as Li⁺ and Na⁺, tend to retain their hydration structure and rarely associate with benzene. Figure 5b shows the number of water molecules in the first solvation shell of Li⁺, Na⁺, K⁺, and NH₄⁺ as a function of the constrained CM separation between the ion and benzene (R_{MX}). These coordination numbers are calculated from the pair distribution function g_{MO} up to the first minimum (2.56, 3.24, and 3.56 Å for Li⁺, Na⁺, and K⁺, respectively,⁴⁶ and 3.37 Å for NH₄⁺).

Figure 5b shows that the presence of a benzene molecule at distances from the ion near the gas phase equilibrium values (1.87, 2.43, 2.83, and 2.92 Å for Li⁺, Na⁺, K⁺, and NH₄⁺, respectively; see Table 1) results in significant loss of ion-coordinated water molecules. The high-energy shoulder of the PMF for lithium corresponds to the deformation of the tetrahedral coordination (at $R_{\text{MX}} < 4.5$ Å) followed by the loss of one of the first-shell water molecules (at $R_{\text{MX}} < 3.0$ Å). The shoulder for sodium corresponds to the loss of the loosely coordinated sixth water molecule from the first hydration shell (at $R_{\text{MX}} < 6$ Å). Upon formation of a complex with benzene, potassium loses 1.6 water molecules, going from a coordination number of 6.9 in bulk water⁴⁶ to 5.3 at $R_{\text{MX}} \sim 3.2$ Å. Ammonium loses only one water molecule, going from a coordination of 5.3 to 4.3 at $R_{\text{MX}} \sim 3.3$ Å. This smaller loss of water molecules (compared to K⁺) is likely the reason why NH₄⁺ associates more strongly and over a longer range than K⁺.

Although the QM and MM complexation energies of Li⁺ and Na⁺ with benzene (see Table 1) are underestimating the experimental enthalpies of formation of the complexes,¹⁰ this does not

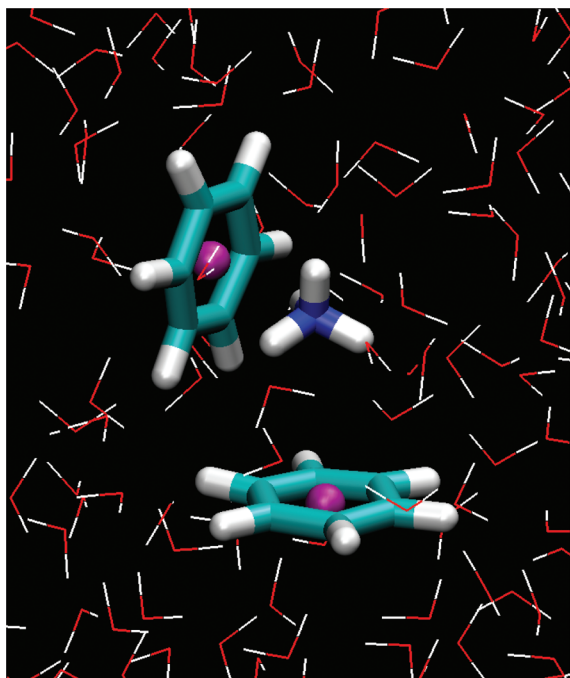


Figure 6. Snapshot of a representative configuration of the NH_4^+ –(benzene)₂ complex in 600 SWM4-NDP water molecules. Atom colors are red for oxygen, blue for nitrogen, cyan for carbon, white for hydrogen, and pink for the nonatomic site of the benzene model.

affect the ion–benzene affinities in aqueous solution. Polarizable models optimized to reproduce the experimental binding energies of the complexes (-39.3 kcal/mol for Li^+ and -22.5 kcal/mol for Na^+) yield PMFs for the lithium–benzene and sodium–benzene pairs in solution that do not deviate from those of Figure 5 by more than 0.1 kcal/mol in the thermodynamically accessible region $R_{\text{MX}} > 3$ Å (data not shown).

Figure 5a also shows the PMF between the centers of two benzene molecules in water, displaying an equilibrium separation of 5.2 Å and a binding free energy of -1.1 kcal/mol, in excellent agreement with the value of -1.00 ± 0.05 kcal/mol for the heat of dimerization of benzene in water reported by Hallén et al.⁶⁶ This binding free energy of the benzene dimer in water represents an improvement over previous simulation results^{67–69} which reported affinities are either too large (-1.5 kcal/mol for ref 67) or too small (-0.5 and -0.36 kcal/mol for refs 68 and 69, respectively).

3.6. Effect of Cations on π – π Interactions in Water. The ab initio calculations reported in Table 1 show that an ion associating to a benzene dimer will form the more stable “ π –cation– π ” sandwich conformations in which two ring systems compete for a direct interaction with the cation, preferably to the less stable “cation– π – π ” conformation in which a cation and a ring system cooperate to bind a central ring system. In that regard, cation– π interactions are disruptive to π – π interactions in the gas phase. To reveal how the interplay between cation– π and π – π interactions translates in aqueous medium, we have performed MD simulations of two benzene molecules in water and two benzene molecules in presence of either one ammonium ion (see Figure 6) or one potassium ion.

The effect of the ion on the association of the two benzene molecules in water can be analyzed from the g_{XX} RDF, where X is benzene center, in presence and absence of the cation (see Figure 7a). The three curves are calculated from 100 ns

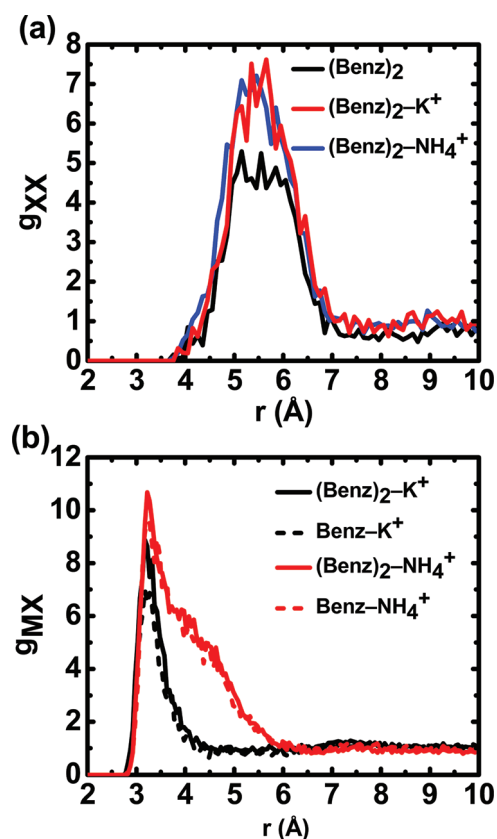


Figure 7. (a) Enhancement of the benzene–benzene radial distribution function (g_{XX}) in the presence of a potassium or an ammonium ion. (b) Enhancement of the ion–benzene radial distribution function (g_{MX}) in the presence of a second benzene molecule. Distributions obtained from 100 ns unconstrained simulations in 600 water molecules.

unconstrained simulations of the binary and ternary complexes and have similar shapes with a broad maximum in the range 5 – 6 Å. The function, however, possesses higher probability for the “ NH_4^+ ” and “ K^+ ” systems, indicating that π – π association increases in presence of the cation.

The influence of the second benzene molecule on the cation– π interaction can similarly be analyzed from the g_{MX} RDF, where M is K^+ or NH_4^+ . Figure 7b shows that the function, calculated from 100 ns unconstrained simulations, has a slightly higher probability for benzene–dimer complexes, compared to the monomer complexes. This indicates that cooperativity exists between cation– π and π – π interactions in aqueous solution.

The most probable arrangement of the two benzene molecules relative to the cation is investigated using 80 ns simulations in which an energy restraint is applied either to prevent the benzene centers of mass from separating by more than 7 Å or to prevent the cation and one benzene molecule from separating by more than 5 Å for K^+ and 6 Å for NH_4^+ . These biased simulations represent the interaction of the ion with a preformed benzene dimer and of the second benzene molecule with a preformed cation–benzene pair. A harmonic force constant of 5 kcal/mol/Å² is used for the restraints.

Figure 8a and c presents the distribution of the ion (K^+ or NH_4^+) as the conditional free energy surface $-k_{\text{B}}T \ln[\rho_{\text{M}}(z_{\text{M}}, r_{\text{M}})/2\pi r_{\text{M}}]$, where $\rho_{\text{M}}(z_{\text{M}}, r_{\text{M}})$ is the ion density relative to the restrained benzene dimer, in cylindrical coordinates. The factor $2\pi r$

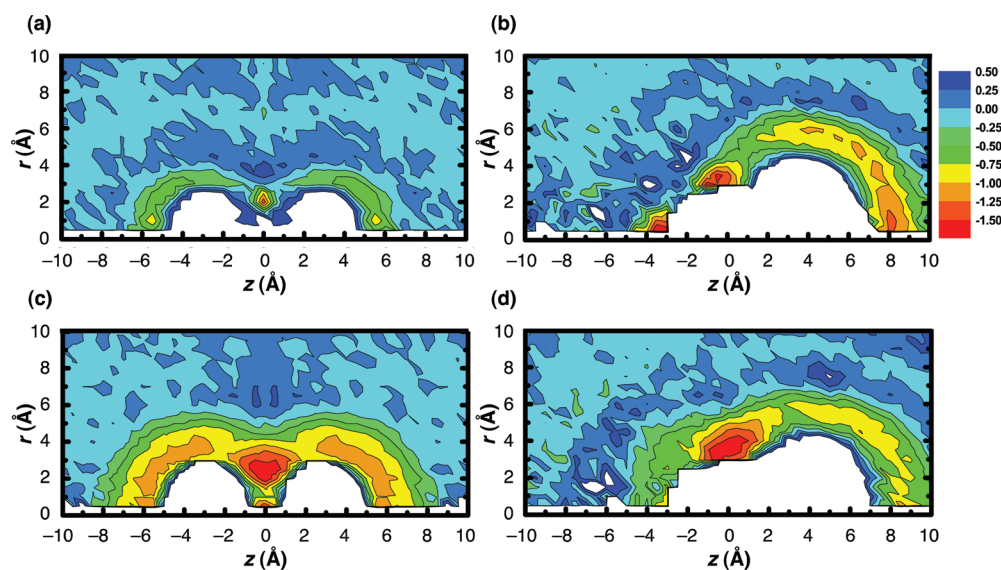


Figure 8. Distribution of (a) K^+ around a preformed benzene dimer, (b) benzene around a preformed K^+ –benzene pair, (c) NH_4^+ around a preformed benzene dimer, and (d) benzene around a preformed NH_4^+ –benzene pair. Densities are presented as the free energy surfaces $-k_B T \ln[\rho(z,r)/2\pi r]$, where $\rho(z,r)$ is the distribution relative to the restrained solutes, in cylindrical coordinates. Benzene molecules are at $z \sim \pm 2.7 \text{ \AA}$ and $r = 0 \text{ \AA}$ in panels (a) and (c). In panels (b) and (d), the ion is at $z = r = 0 \text{ \AA}$ and the benzene molecule is at $z \sim 3.2 \text{ \AA}$ and $r = 0 \text{ \AA}$.

accounts for the purely geometric probability of finding the ion at a radial distance r from the benzene–benzene axis. Figure 8b and d shows function $-k_B T \ln[\rho_X(z_X, r_X)/2\pi r_X]$, the conditional free energy surface for a benzene molecule in the presence of a restrained ion–benzene dimer.

Figure 8a shows a high-density K^+ “envelope” at a distance of 3.0–3.5 \AA from either one of the two benzene molecules (which are located at $z \sim \pm 2.6 \text{ \AA}$ and $r = 0 \text{ \AA}$). This region corresponds to the minimum of the K^+ –benzene PMF of Figure 5a. Maximum ion density is found at $z = 0 \text{ \AA}$ and $r \sim 2.0 \text{ \AA}$, where the ion is coordinating both benzene molecules and forming an “isosceles triangle” conformation, and at $z \sim 5.5 \text{ \AA}$ and $r < 2 \text{ \AA}$, where the system is forming a K^+ –benzene–benzene stacked conformation.

Figure 8b shows a high-density benzene envelope at a distance of 5–6 \AA from the ion-bound benzene molecule (located at $z \sim 3.2 \text{ \AA}$ and $r = 0 \text{ \AA}$) or at a distance of 3.0–3.5 \AA from the K^+ ion (located at $z = r = 0 \text{ \AA}$). Three regions are markedly populated: the “triangle” conformation at $z \sim 0 \text{ \AA}$ and $r \sim 3.5 \text{ \AA}$, the K^+ –benzene–benzene stacked conformation at $z \sim 8 \text{ \AA}$ and $r < 2 \text{ \AA}$, and the benzene– K^+ –benzene sandwich conformation at $z \sim -3.5 \text{ \AA}$ and $r < 1 \text{ \AA}$. This sandwich conformation is not observed in Figure 8a due to the weak restraint on the benzene molecules, preventing them from separating. The “triangle” conformation is the most favorable arrangement and corresponds to a potassium ion coordinated by two benzene molecules forming an $X \cdots M \cdots X$ angle of $\sim 90^\circ$. Interestingly, the bent sandwich conformation, located between the “triangle” and sandwich conformations of Figure 8b, is poorly populated despite being the most stable structure in gas phase (see Table 1). This may be attributed to the higher degree of dehydration of the K^+ ion accompanying this arrangement. Indeed, three additional simulations of the ternary complex with restraints keeping the $X \cdots M$ distances within 3.2 \AA and the $X \cdots M \cdots X$ angle around 90° , 120° , and 180° show that the K^+ ion is coordinated by 4.1 water molecules for $X \cdots M \cdots X \sim 90^\circ$, 3.7 molecules for $X \cdots M \cdots X \sim 120^\circ$, and 4.1 molecules for $X \cdots M \cdots X \sim 180^\circ$. This shows that K^+ is best hydrated when the complex adopts the “triangle” geometry

($X \cdots M \cdots X \sim 90^\circ$) and poorly hydrated when it adopts the bent sandwich geometry ($X \cdots M \cdots X \sim 120^\circ$). It also confirms that the “triangle” and straight sandwich arrangements are equally favorable, as both provide a solvation shell formed of two benzene and four water molecules.

The simulations involving NH_4^+ display similar features (see Figure 8c and d) but with a marked preference for the triangle arrangement. The stacked and sandwiched geometries do not appear particularly favored or disfavored. Additional simulations with restraints keeping the $X \cdots M$ distances below 3.3 \AA , and the $X \cdots M \cdots X$ angle at 90° , 120° , and 180° shows that the NH_4^+ ion is coordinated by 3.6 water molecules for $X \cdots M \cdots X \sim 90^\circ$ and 3.3 water molecules for $X \cdots M \cdots X$ angles of 120° and 180° . This confirms that the “triangle” conformation is the most favored and that straight and bent sandwich conformers are equally accessible.

4. CONCLUSION

Ab initio calculations on cation– π interactions of Li^+ , Na^+ , K^+ , and NH_4^+ with benzene show that cation– π interactions in the gas phase become stronger if additional aromatic systems are introduced in stacked arrangements but weaker if they are introduced in sandwiched arrangements. The dominant contribution to this cooperativity effect is electronic polarization. Polarizable models are parametrized to reproduce the interactions of Li^+ , Na^+ , K^+ , and NH_4^+ with benzene as well as the interaction of NH_4^+ with water. The models reproduce the ab initio energies of both the benzene complexes and the water complexes. The model for NH_4^+ also reproduces the experimental hydration free energy of the ion without further adjustment.

Potentials of mean force between cation–benzene pairs in aqueous solution show that, while Li^+ and Na^+ do not bind benzene, K^+ and NH_4^+ bind with free energies of -1.2 and -1.4 kcal/mol, respectively. The small Li^+ and Na^+ ions form rigid complexes with their coordinating water molecules that cannot easily accommodate the replacement of a water molecule by a benzene molecule. The

larger K^+ and NH_4^+ ions, on the other hand, have more flexible hydration shells that are favorable to benzene inclusion.

Methyl substitution of ammonium hydrogens is reported to decrease the gas phase binding energy between the ions and benzene,^{19,70} For example, while we report a binding energy of -17.58 kcal/mol between ammonium and benzene, Xu et al.¹⁹ have reported a binding energy of -15.78 kcal/mol between methylammonium and benzene at the same level of theory and Felder et al.⁷⁰ have reported a binding energy of -8.40 kcal/mol between tetramethylammonium and benzene at the MP2/6-31G(d) level of theory. It would be interesting to use similarly derived polarizable models to investigate the enhanced hydrophobic association with benzene resulting from successive methylations of the ammonium ion.

The results provide new insight on the influence of inorganic salts on the solubility of aromatic hydrocarbons in water, the so-called “salting-out” effect.^{71,72} The degree of salting-out for aromatic compounds depends on the salt in an apparently nonsystematic way. For benzene, the salting-out effect in presence of lithium, sodium, potassium, and ammonium chloride salts follows the order $Na^+ > K^+ > Li^+ > NH_4^+$.⁷¹ Based on the binding free energies calculated in this work, we suggest that the salting-out effect follows two distinct mechanisms. Li^+ and Na^+ , which expel benzene molecules from their first hydration shells, are effectively decreasing the volume of solution available to benzene— Na^+ more so than Li^+ —, leading to benzene association and salting-out.⁷³ K^+ and NH_4^+ are also effectively decreasing the volume of solution available to benzene but according not only to their sizes (which are comparable) but also to their affinities for benzene. Since NH_4^+ has a higher (and longer-range) affinity for benzene than K^+ , the excluded volume it creates is smaller, and its salting-out effect is expected to be weaker. It is important to note, however, that the present explanation does not account for the counterion (e.g., Cl^-), which might play different roles in presence of different cations, as it was suggested for lithium chloride.⁷⁴

While K^+ —(benzene)₂ and NH_4^+ —(benzene)₂ complexes in gas phase are significantly more stable in their sandwiched conformations than in their stacked conformations, the reverse is true in aqueous solution. The most stable arrangement of the two systems in water, however, corresponds to a “triangle” geometry in which the two benzene molecules are both directly coordinating the cation, at an angle of approximately 90° . This geometry preserves the benzene—benzene hydrophobic interaction, while minimizing ion dehydration.

The concepts put forward in the present work can be transposed to cation— π interactions in proteins. While a single aromatic residue, such as phenylalanine or tryptophan, at the surface of a protein may not create enough binding affinity for K^+ or NH_4^+ ions to generate significant biological function, a binding site formed of multiple aromatic residues may.^{75–77} Our simulations suggest that an L-shaped arrangement of two aromatic residues binds ions better than a stacked or sandwiched arrangement. This thermodynamic advantage may however not be a dominant driving factor in light of the fact that, unlike in solution, residues forming “aromatic cages” in proteins are likely to be preassembled and optimally oriented. Nevertheless, the fact that K^+ and NH_4^+ ions have different affinities for stacked and sandwiched arrangements of aromatic residues can possibly be exploited to design protein receptors selective to potassium or ammonium that rely more on selectively accommodating a partially hydrated ion than on providing a selective full coordination environment.⁷⁸

The parametrization procedure introduced in this work can be readily applied to other aromatic moieties to provide a set of polarizable models representing the various cation— π interactions found in proteins. Such models can likely help to elucidate the factors determining the abundances of the various cation— π and cation— π_2 conformations surveyed in the Protein Data Bank^{79,80} and can be used to study biological systems in which cation— π interactions play important roles.

AUTHOR INFORMATION

Corresponding Author

*E-mail: guillaume.lamoureux@concordia.ca

†On leave from Department of Chemistry, Faculty of Science, Assiut University, Assiut 71516, Egypt.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

This work was supported by an FQRNT Établissement de nouveaux chercheurs grant (NC-125413) and by a PROTEO scholarship to EAO. Computational resources were provided by the Réseau québécois de calcul de haute performance (RQCHP) and by an FQRNT Equipment grant.

REFERENCES

- (1) Dougherty, D. A. *Science* **1996**, *271*, 163–168.
- (2) Mecozzi, S.; West, A. P.; Dougherty, D. A. *J. Am. Chem. Soc.* **1996**, *118*, 2307–2308.
- (3) Mecozzi, S.; West, A. P.; Dougherty, D. A. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 10566–10571.
- (4) Ma, J. C.; Dougherty, D. A. *Chem. Rev.* **1997**, *97*, 1303–1324.
- (5) Gallivan, J. P.; Dougherty, D. A. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9459–9464.
- (6) Woodin, R. L.; Beauchamp, J. L. *J. Am. Chem. Soc.* **1978**, *100*, 501–507.
- (7) Deakynne, C. A.; Meotner, M. *J. Am. Chem. Soc.* **1985**, *107*, 474–479.
- (8) Wouters, J. *Protein Sci.* **1998**, *7*, 2472–2475.
- (9) Armentrout, P. B.; Rodgers, M. T. *J. Phys. Chem. A* **2000**, *104*, 2238–2247.
- (10) Amicangelo, J. C.; Armentrout, P. B. *J. Phys. Chem. A* **2000**, *104*, 11420–11432.
- (11) Amunugama, R.; Rodgers, M. T. *J. Phys. Chem. A* **2002**, *106*, 5529–5539.
- (12) Ruan, C. H.; Rodgers, M. T. *J. Am. Chem. Soc.* **2004**, *126*, 14600–14610.
- (13) Yorita, H.; Otomo, K.; Hiramatsu, H.; Toyama, A.; Miura, T.; Takeuchi, H. *J. Am. Chem. Soc.* **2008**, *130*, 15266–15267.
- (14) Hallowita, N.; Carl, D. R.; Armentrout, P. B.; Rodgers, M. T. *J. Phys. Chem. A* **2008**, *112*, 7996–8008.
- (15) Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 4177–4178.
- (16) Cubero, E.; Luque, F. J.; Orozco, M. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 5976–5980.
- (17) Feller, D.; Dixon, D. A.; Nicholas, J. B. *J. Phys. Chem. A* **2000**, *104*, 11414–11419.
- (18) Tsuzuki, S.; Yoshida, M.; Uchamaru, T.; Mikami, M. *J. Phys. Chem. A* **2001**, *105*, 769–773.
- (19) Xu, Y.; Shen, J.; Zhu, W.; Luo, X.; Chen, K.; Jiang, H. *J. Phys. Chem. B* **2005**, *109*, 5945–5949.
- (20) Coletti, C.; Re, N. *J. Phys. Chem. A* **2006**, *110*, 6563–6570.

- (21) Frontera, A.; Quinonero, D.; Garau, C.; Costa, A.; Ballester, P.; Deya, P. M. *J. Phys. Chem. A* **2006**, *110*, 9307–9309.
- (22) Soteras, I.; Orozco, M.; Luque, F. J. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2616–2624.
- (23) Mishra, B. K.; Bajpai, V. K.; Ramanathan, V.; Gadre, S. R.; Sathyamurthy, N. *Mol. Phys.* **2008**, *106*, 1557–1566.
- (24) Marshall, M. S.; Steele, R. P.; Thanthirivatt, K. S.; Sherrill, C. D. *J. Phys. Chem. A* **2009**, *113*, 13628–13632.
- (25) Vijay, D.; Sastry, G. N. *Chem. Phys. Lett.* **2010**, *485*, 235–242.
- (26) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. *J. Mol. Biol.* **1977**, *112*, 535–542.
- (27) Minoux, H.; Chipot, C. *J. Am. Chem. Soc.* **1999**, *121*, 10366–10372.
- (28) Crowley, P. B.; Golovin, A. *Proteins: Struct. Funct. Bioinf.* **2005**, *59*, 231–239.
- (29) Wintjens, R.; Lievin, J.; Rooman, M.; Buisine, E. *J. Mol. Biol.* **2000**, *302*, 395–410.
- (30) Gromiha, M. M.; Santhosh, C.; Ahmad, S. *Int. J. Biol. Macromol.* **2004**, *34*, 203–211.
- (31) Shi, Z. S.; Olson, C. A.; Kallenbach, N. R. *J. Am. Chem. Soc.* **2002**, *124*, 3284–3291.
- (32) Prajapati, R. S.; Sirajuddin, M.; Durani, V.; Sreeramulu, S.; Varadarajan, R. *Biochemistry* **2006**, *45*, 15000–15010.
- (33) Lummis, S. C. R.; Beene, D. L.; Harrison, N. J.; Lester, H. A.; Dougherty, D. A. *Chem. Biol.* **2005**, *12*, 993–997.
- (34) Lehn, J.-M.; Meric, R.; Vigneron, J.-P.; Cesario, M.; Guilhem, J.; Pascard, C.; Asfari, Z.; Vicens, J. *Supramol. Chem.* **1995**, *5*, 97–103.
- (35) Archambault, F.; Chipot, C.; Soteras, I.; Luque, F. J.; Schulten, K.; Dehez, F. *J. Chem. Theory Comput.* **2009**, *5*, 3022–3031.
- (36) Sunner, J.; Nishizawa, K.; Kebarle, P. *J. Phys. Chem.* **1981**, *85*, 1814–1820.
- (37) Chipot, C.; Maigret, B.; Pearlman, D. A.; Kollman, P. A. *J. Am. Chem. Soc.* **1996**, *118*, 2998–3005.
- (38) Costanzo, F.; Della Valle, R. G.; Barone, V. *J. Phys. Chem. B* **2005**, *109*, 23016–23023.
- (39) Sa, R. J.; Zhu, W. L.; Shen, J. H.; Gong, Z.; Cheng, J. G.; Chen, K. X.; Jiang, H. L. *J. Phys. Chem. B* **2006**, *110*, 5094–5098.
- (40) (a) Drude, P. *Lehrbuch der Optik*; S. Hirzel: Leipzig, Germany, 1900. English translation: *The Theory of Optics*; Drude, P., translated from the German by Riborg Mann, C.; Millikan, R. A.; Dover Publications, Inc.: Mineola, New York, 1907.
- (41) Lamoureux, G.; Roux, B. *J. Chem. Phys.* **2003**, *119*, 3025–3039.
- (42) Anisimov, V. M.; Lamoureux, G.; Vorobyov, I. V.; Huang, N.; Roux, B.; MacKerell, A. D. *J. Chem. Theory Comput.* **2005**, *1*, 153–168.
- (43) Frisch, M. J. T.; G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. *J. Gaussian 09*, revision B.01; Gaussian, Inc., Wallingford, CT, 2009.
- (44) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553–566.
- (45) Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caffisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. *J. Comput. Chem.* **2009**, *30*, 1545–1614.
- (46) Yu, H. B.; Whitfield, T. W.; Harder, E.; Lamoureux, G.; Vorobyov, I.; Anisimov, V. M.; MacKerell, A. D.; Roux, B. *J. Chem. Theory Comput.* **2010**, *6*, 774–786.
- (47) 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.
- (48) Lamoureux, G.; Roux, B. *J. Phys. Chem. B* **2006**, *110*, 3308–3322.
- (49) Lopes, P. E. M.; Lamoureux, G.; Roux, B.; MacKerell, A. D. *J. Phys. Chem. B* **2007**, *111*, 2873–2885.
- (50) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.
- (51) Lamoureux, G.; Harder, E.; Vorobyov, I. V.; Roux, B.; MacKerell, A. D. *Chem. Phys. Lett.* **2006**, *418*, 245–249.
- (52) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593.
- (53) Lagüe, P.; Pastor, R. W.; Brooks, B. R. *J. Phys. Chem. B* **2004**, *108*, 363–368.
- (54) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011–1021.
- (55) Souaille, M.; Roux, B. *Comput. Phys. Commun.* **2001**, *135*, 40–57.
- (56) Khavrutskii, I. V.; Dzubiella, J.; McCammon, J. A. *J. Chem. Phys.* **2008**, *128*, 044106.
- (57) Brugé, F.; Bernasconi, M.; Parrinello, M. *J. Chem. Phys.* **1999**, *110*, 4734–4736.
- (58) Brugé, F.; Bernasconi, M.; Parrinello, M. *J. Am. Chem. Soc.* **1999**, *121*, 10883–10888.
- (59) Klots, C. E. *J. Phys. Chem.* **1981**, *85*, 3585–3588.
- (60) Contreras, R.; Klopman, G. *Can. J. Chem.* **1985**, *63*, 1746–1749.
- (61) Åqvist, J. *J. Phys. Chem.* **1990**, *94*, 8021–8024.
- (62) Marcus, Y. *J. Chem. Soc., Faraday Trans* **1991**, *87*, 2995–2999.
- (63) Tissandier, M. D.; Cowen, K. A.; Feng, W. Y.; Gundlach, E.; Cohen, M. H.; Earhart, A. D.; Coe, J. V.; Tuttle, T. R. *J. Phys. Chem. A* **1998**, *102*, 7787–7794.
- (64) Ben-Naim, A.; Marcus, Y. *J. Chem. Phys.* **1984**, *81*, 2016–2027.
- (65) Kumpf, R. A.; Dougherty, D. A. *Science* **1993**, *261*, 1708–1710.
- (66) Hallén, D.; Wadsö, I.; Wasserman, D. J.; Robert, C. H.; Gill, S. J. *J. Phys. Chem.* **1988**, *92*, 3623–3625.
- (67) Jorgensen, W. L.; Severance, D. L. *J. Am. Chem. Soc.* **1990**, *112*, 4768–4774.
- (68) Linse, P. *J. Am. Chem. Soc.* **1993**, *115*, 8793–8797.
- (69) Chipot, C.; Jaffe, R.; Maigret, B.; Pearlman, D. A.; Kollman, P. A. *J. Am. Chem. Soc.* **1996**, *118*, 11217–11224.
- (70) Felder, C.; Jiang, H.-L.; Zhu, W.-L.; Chen, K.-X.; Silman, I.; Botti, S. A.; Sussman, J. L. *J. Phys. Chem. A* **2001**, *105*, 1326–1333.
- (71) McDevit, W. F.; Long, F. A. *J. Am. Chem. Soc.* **1952**, *74*, 1773–1777.
- (72) Sanemasa, I.; Arakawa, S.; Araki, M.; Deguchi, D. *Bull. Chem. Soc. Jpn.* **1984**, *57*, 1539–1544.
- (73) Kalra, A.; Tugcu, N.; Cramer, S. M.; Garde, S. *J. Phys. Chem. B* **2001**, *105*, 6380–6386.
- (74) Thomas, A. S.; Elcock, A. H. *J. Am. Chem. Soc.* **2007**, *129*, 14887–14898.
- (75) Campagna-Slater, V.; Schapira, M. *Mol. Inf.* **2010**, *29*, 322–331.
- (76) Khademi, S.; O’Connell, J.; Remis, J.; Robles-Colmenares, Y.; Miercke, L. J.; Stroud, R. M. *Science* **2004**, *305*, 1587–1594.
- (77) Zheng, L.; Kostrewa, D.; Bernèche, S.; Winkler, F. K.; Li, X. D. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 17090–17095.
- (78) Chin, J.; Walsdorff, C.; Stranix, B.; Oh, J.; Chung, H. J.; Park, S.-M.; Kim, K. *Angew. Chem., Int. Ed.* **1999**, *38*, 2756–2759.
- (79) Reddy, A. S.; Vijay, D.; Sastry, G. M.; Sastry, G. N. *J. Phys. Chem. B* **2006**, *110*, 2479–2481.
- (80) Chelli, R.; Procacci, P. *J. Phys. Chem. B* **2006**, *110*, 10204–10205.

A Multi-Objective Approach to Force Field Optimization: Structures and Spin State Energetics of d^6 Fe(II) Complexes

Christopher M. Handley and Robert J. Deeth*

Inorganic Computational Chemistry Group, Department of Chemistry, Univ. of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, Great Britain

S Supporting Information

ABSTRACT: The next generation of force fields (FFs), regardless of the accuracy of the potential energy representation, will always have parameters that must be fitted in order to reproduce experimental and/or *ab initio* data accurately. Single objective methods have been used for many years to automate the obtaining of parameters, but this leads to ambiguity. The solution depends on the chosen weights and is therefore not unique. There have been few advances in solving this problem, which thus remains a major hurdle for the development of empirical FF methods. We propose a solution based on multi-objective evolutionary algorithms (MOEAs). MOEAs allow the FF to be tuned against the desired objectives and offer a powerful, efficient, and automated means to reparameterize FFs, or even discover the parameters for a new potential. Here, we illustrate the application of MOEAs by reparameterizing the ligand field molecular mechanics (LFMM) FF recently reported for modeling spin crossover in iron(II)–amine complexes (Deeth et al. *J. Am. Chem. Soc.* **2010**, *132*, 6876). We quickly recover the performance of the original parameter set and then significantly improve it to reproduce the geometries and spin state energy differences of an extended series of complexes with RMSD errors in Fe–N and N–N distances reduced from 0.06 Å to 0.03 Å and spin state energy difference RMSDs reduced from 1.5 kcal mol⁻¹ to 0.2 kcal mol⁻¹. The new parameter sets highlight, and help resolve, shortcomings both in the non-LFMM FF parameters and in the interpretation of experimental data for several other Fe(II)N₆ amine complexes not used in the FF optimization.

INTRODUCTION

Molecular modeling and simulation are powerful tools for probing many biological and chemical processes. Treating systems comprising many thousands or even millions of particles has become more or less routine, yet despite the power of contemporary computers, simulations using quantum mechanical (QM) methods (e.g., density functional theory (DFT)) remain relatively scarce and restricted to picosecond time scales. Hence, the development of quantitatively accurate empirical force fields (FFs) is still a high priority.

Conventional molecular mechanics (MM) techniques have been well developed for simulating systems that consist of many light atoms.^{1–4} However, introducing heavy atoms, specifically transition metals (TM), into simulations presents additional complications due to their “electronic activity”. TMs typically have partially filled d orbitals and support a variety of oxidation states, either of which may significantly affect the stability and structure of a given complex.

These issues, and how to deal with them within MM simulations, including our own ligand field molecular mechanics (LFMM) approach,⁵ have been proposed and reviewed.^{6,7} This paper deals with the even more fundamental issue of how to derive the FF parameters in the first place.

As with all FF approaches, parametrization is critical but is often considered more of a black art than a science.^{8–10} The problem is exacerbated when introducing TMs into MM simulations since the additional functions, and their attendant parameters, required to describe the metal–ligand interactions introduce additional complexity over “conventional” MM.¹¹

Ideally, therefore, we seek automated processes that can find the optimal LFMM parameters.

There are two problems in parameter fitting: the choice of the training data and the fitting of the parameters to these training data. Typically, most FFs are parametrized to fit *ab initio* or experimental data. The exact nature of the fit depends upon the objectives that must be minimized. Ideally, the FF should reproduce accurate relative conformational energies for the complexes, as well as accurate forces and configurations (i.e., geometries) at the minima. Most parametrization methods are automated and focus on simultaneously optimizing a number of quantities (e.g., energies, forces, Hessian matrix elements, and configurations) by minimizing a single penalty function, such as the sum of the least-squares deviation.^{12–15} Other automated parametrization methods have used gradients,^{11,12,16,17} neural networks,^{18–21} genetic programming,²² and genetic algorithms.^{8,23–27} However, the approach used by Mostaghim et al.²³ is the only one to use a multi-objective optimization algorithm (MOOA).²⁸

MOOAs promise a step change in automatic parametrization which could revolutionize the application of FF methods to increasingly complex molecular and solid-state systems. This work describes our implementation of a MOOA and its first application within the LFMM framework. To illustrate the new approach, we reanalyze the MMFF94/LFMM force field for d^6 iron(II) amine spin complexes.²⁹ Six-coordinate iron(II) complexes support two spin states and, given the right ligand set,

Received: August 20, 2011

Published: December 01, 2011

display spin crossover (SCO) behavior. SCO represents both a challenge for molecular modeling and a significant target for functional materials which can be used as molecular switches. We demonstrate that the MOOA approach rapidly recovers the original parameter set and locates a number of improved sets which are then applied to other Fe complexes. The improved parameters reveal outlier complexes which contain structural elements in the ligands which were not covered by the original training systems. In particular, cage complexes possess six-membered chelate rings which are forced by the ligand structure to adopt boat conformations which differ for high spin and low spin versions and require modification of some of the (non-LFMM) torsional potentials. The MOOA can then be redeployed to rapidly reoptimize the LFMM parameters and regenerate a balanced FF.

MULTI-OBJECTIVE OPTIMIZATION ALGORITHMS

Most optimization methods make use of gradient descent algorithms where the aim is to find the parameters that correspond to a minimum on an error surface. The error, or objective, which is being minimized, is often the sum of the weighted sum of squared errors for a number of properties within the training set. The weighting of the sum-of-squared errors is required for two reasons. First, we can adjust the weights so that a particular property within the training set is fitted more closely. But also the weighting must be correctly allocated so that the influence of different properties is a fair reflection of the different nature of the errors (for example the difference between bond distance errors and angle variations).

Weighting of each sum-of-property error leads to a penalty function which can be minimized. However, the final set of parameters is only valid for the weights that have been allocated in the penalty function and is the only one found when in fact there can be many parameter sets that are all valid solutions. In order for there to be no ambiguity in the weights selected, prior knowledge about the force field, and the influence of the weights on the predictive power for each of the properties, is required.

The alternative to the single objective approach is to use MOOA where a number of objectives is optimized at once. There is no one solution that is optimal; rather there are a number of solutions that are all optimal. With respect to the search space, these solutions are better than all others when all objectives are considered. Thus, this set of solutions is the *Pareto-optimal* set.³⁰

As an example, we can consider the optimal design of a super-computer that can perform many calculations but that must also be energy efficient. These objectives are competing. The computer can be more powerful, but in turn consume more power. If it consumes less power, it may well not be as powerful a computer. None of the Pareto-optimal solutions is superior to any other, but they all are superior to the (nonoptimal) solutions. MOOA allows for a search of the design space and can aid in the design-making process.

Within MOOA approaches, genetic algorithms (GAs) are the most popular, while Kriging has also been applied to the MOOAs.³¹ The use of GAs within MOOA gives us the term multi-objective evolutionary algorithm (MOEA).^{30,32,33} The popularity of GAs and other evolutionary algorithms (EAs) exists because they process an entire population of solutions with each generation. The EA can then be used to generate new populations of solutions, while retaining strong parameters but promoting diversity.

In this way, an EA can discover many solutions that are members of the Pareto-optimal set of solutions.

Within the field of chemistry, MOOAs and MOEAs have seen some use, particularly with respect to drug discovery, as reviewed by Nicolaou et al.²⁸ Force field design has had only limited exposure to these new techniques.^{23,34}

The main point of MOEAs is that the Pareto front of solutions dominates all other solutions found with respect to all objectives. Within the Pareto front set of solutions, each solution dominates another solution in all objectives but one. This represents the fact that there can be no improvement in one objective without loss of performance in another.

Within MOEA, we can be more precise. We wish to minimize \vec{y} , the objective vector.

$$\vec{y} = \vec{f}(\vec{x}) = (f_1(\vec{x}), f_2(\vec{x}), \dots, f_m(\vec{x})) \quad (1)$$

Here, the objective vector is a function of the decision vector (parameters) $\vec{x} = (x_1, x_2, \dots, x_n)^T$, and the decision vectors reside in the parameter search space of $S \subset \mathcal{R}^n$. The image (performance in the objectives) of the decision vectors lies within the objective space $Z = \subset \mathcal{R}^m$. Thus, the elements of Z are the objective vectors and are composed of the objective values, $\vec{f}(\vec{x}) = (f_1(\vec{x}), f_2(\vec{x}), \dots, f_m(\vec{x}))$.²³ \mathcal{R}^m and \mathcal{R}^n are m and n dimensional real number spaces.

A decision vector (set of parameters), x_1 , *dominates* another decision vector, x_2 , if x_1 is not any worse than x_2 in any objective while still outperforming x_2 in one objective.

If x_1 is not worse than x_2 in any of the objectives, though not better than x_2 , then x_1 and x_2 are deemed *equivalent*. Solutions are also equivalent if they are worse in one objective or more, while being better in others.

If x_1 is not dominated by any solution then it is deemed *Pareto Optimal*. This means all members of the Pareto Optimal set are better than each other only in one objective. Members of the Pareto Optimal set in the decision space form the Pareto Optimal Front in the image space.

In this paper, the aim is to identify the Pareto Optimal sets of parameters for the LFMM force field, where we are trying to optimize two objectives (an energy error and a distance error).

Finding these Pareto Optimal sets is necessary for further generations of the EA. A GA takes a population of known solutions and uses them to create a new population of solutions. This is performed by the simple act of crossover and mutation of the bit strings that represent the decision vector. The members of the Pareto Optimal set can be used to guide the generation of further solutions by maintaining the best strengths of previous Pareto Optimal solutions. There are a number of MOEA approaches, but here we make use of nondominated sorting genetic algorithm II (NSGA-II).³⁵

Ultimately, the use of MOOAs allows for an efficient exploration of parameters, removing the need for the user to have in depth prior knowledge about the target system or the potentials used. Instead, the user is simply required to choose, from a selection of Pareto Optimal solutions, the parameter set/s that best serves their needs.

METHOD

We follow the previous work of Deeth et al.²⁹ on iron amine complexes for the generation of the *ab initio* training data. All DFT calculations use the Amsterdam Density Functional program version 2008.01.³⁶ As suggested by Swart,³⁷ all calculations

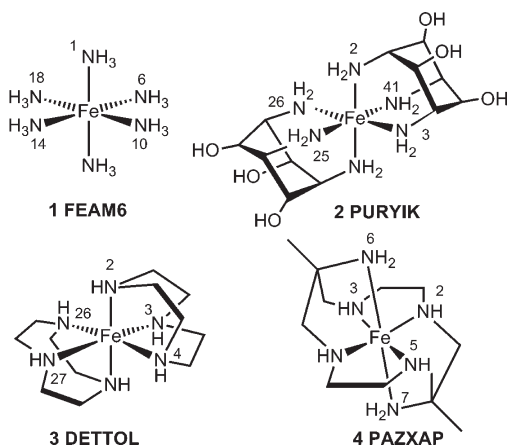


Figure 1. Schematic structures and numbering schemes for complexes used in initial LFMM FF optimization.

were performed using the OPBE functional, uncontracted triple- ζ plus polarization STO basis sets (TZP). In addition, we employ the conductor-like screening model (COSMO) to include the influence of condensed-phase effects ($\epsilon = 78$, probe radius = 1.9 Å).³⁸ (Cartesian coordinates and ADF total binding energies for all the iron amine complexes are given in Table S1 of the Supporting Information.)

For our proof-of-principle test, we began with the original training set²⁹ (Figure 1). The data for this set include the optimized structures and energies for the high spin ($S = 2$) and low spin ($S = 0$) versions of four d⁶ Fe(II) complexes: [Fe(NH₃)₆]²⁺ (FEAM6), [Fe(tachOH)₂]²⁺ (PURYIK, tachOH = 1,3,5-triamino-2,4,6-trihydrocyclohexane³⁹), [Fe([9]aneN₃)₂]²⁺ (DETTOL, [9]aneN₃ = 1,4,7-triazacyclononane³⁹), and [Fe(diammac)]²⁺ (PAZXAP, diammac = *exo*-6,13-diamino-6,13-dimehyl-1,4,8,11-tetraazatetradecane). These complexes that span the SCO divide have been previously studied with respect to the design of potential spin crossover and light induced excited spin state trapping (LIESST) complexes.²⁹ Optimization and testing of the new force field parameters occurs within the DommiMOE⁴⁰ program as implemented in MOE 2010.⁴¹ The new routine, PROTEUS (PaReto OpTimal EvolUtionary System), works with DommiMOE and optimizes the LFMM parameters for the interactions that the user has selected. We will now review the NSGA-II algorithm and detail how it has been implemented for parameter optimization.

In order to initiate parameter optimization, we first require a set of parameters. This would normally be a user's first best guess which acts as the seed from which PROTEUS explores parameter space. For LFMM, we have a number of parameters to optimize which must be varied to differing degrees. These arise from the Morse function for M–L stretches, ligand–ligand repulsion parameters which help define the angular geometry of the complexes, and angular overlap model (AOM) parameters which determine the d-orbital energies and thus the ligand field stabilization energy (LFSE).⁴⁰ Within the Morse functions, the reference distances, r_0 , are varied by up to 1.5 Å in either direction, the dissociation energies, D , by up to 75% in either direction, and the α coefficients by up to 75% in either direction. Within the ligand–ligand repulsion terms, the A_{LL} parameter changes by up to 5%, while the power to which the term is raised is kept constant. The AOM parameters, e_{aom} , are derived from the $a_{\text{aom},n}$ where aom refers to the symmetry of the M–L interaction

(i.e., σ or π) and n refers to the power to which the distance between atoms is raised. In general, $e_{\text{aom}} = a_{\text{aom},n}/r_{\text{M-L}}^n$. The $a_{\text{aom},n}$ parameters were allowed to vary by up to 2000 units in either direction. For example, for our systems, $a_{\sigma 6}$ starts at a value of 41 300. This means we vary this parameter by only 0.5%. But for the electron pairing energy parameters, the variation is much larger due to the smaller magnitude of these parameters at the start of training. These limits of variation can later be reduced (or enlarged) to allow for a finer (or coarser) grade search of the local parameter space, as and when required.

Using these limits of variation of the parameters, we generate 10N sets of parameters, where N is the number of parameters that are being optimized. Each parameter set is determined randomly by generating a string of bits. Each bit is randomly determined when the string is generated. This long string of bits can then be sequentially broken up into N substrings, each of 30 bits. Later, we increase the size of these substrings to allow for a fine grade search of the local parameter space, since a longer substring length allows for smaller variations in the parameters. These substrings of bits are then decoded into real values that each lie within the limits of variation allowed for each parameter. Therefore, the initial bit string, known as a chromosome, consists of a string of zeros and ones that is NM long, where M is the length of each individual substring, in our initial case, 30.

Once each chromosome has been generated, they are decoded to give a new parameter set for which the fitness functions are evaluated. The fitness functions we use are the RMSDs between DFT and LFMM spin-state energy differences, and the RMSDs for Fe–N and N–N distances. The aim is to minimize these fitness functions.

When each parameter set has been used and the fitness functions have been evaluated for each set, the population can be sorted into Pareto fronts (following the rules outlined earlier) and assigned a rank wherein a given solution dominates all others of lower rank. The solutions are also assigned a density which describes how dense the solutions are about a particular solution. The density is the sum of distances in each of the objectives between the two nearest solutions. Using both the Pareto rank of a solution and the density, we can create new sets of parameters.

A new parameter set is created by *mating* and *mutation*. For mating, two different parameter set chromosomes are crossed over at a randomly determined point along the chromosome. This means that all bits beyond the selected chromosome now are equal to those found in the other chromosome, and vice versa. Once mating has happened, the two new *children* are mutated. Each bit in each new chromosome may be randomly changed to its opposite value.

Mating and mutation does not occur for every member of the original *parent* population. Instead, parents are selected on the basis of their Pareto rank and the density of solutions about them. For a new population of Q solutions, Q parents must be selected, as each pair of parents generates two children solutions. The parents are selected by tournament. For each parent chosen, two are randomly selected from the entire population. The rank is compared, and we retain the potential parent with the lowest rank. If the ranks are the same, then the parent with the lowest density is chosen. This method ensures elitism—i.e., the best parents should give rise to the best children—and diversity—i.e., it ensures that the parents are not too similar, which would give rise to children that do not adequately explore the parameter space.

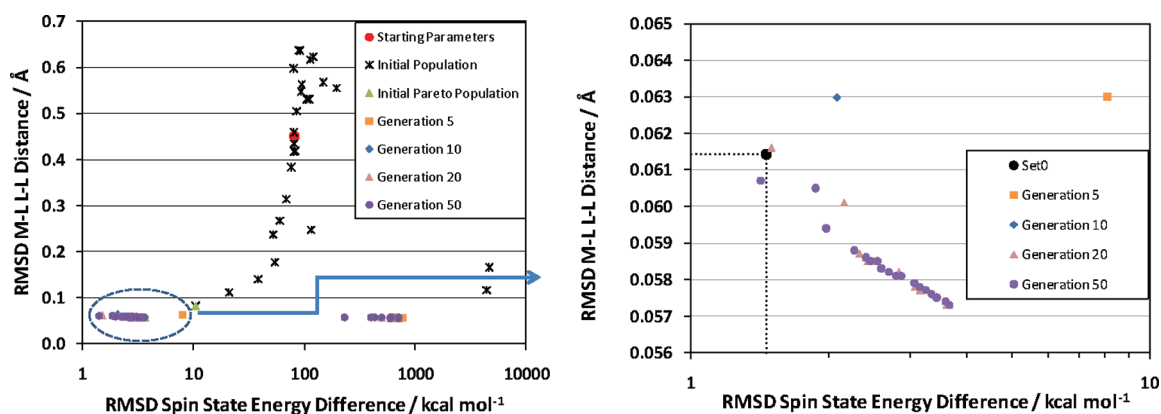


Figure 2. Left: The initial population of solutions (crosses) generated from the starting parameters (red dot on left-hand plot) is refined in round 1 of optimization to a Pareto front from which a point is selected (green triangle, the other points are not shown) to initiate round 2 of optimization: fifth generation (orange squares), 10th generation (blue diamonds), 20th generation (pink triangles), and 50th generation (purple dots). Right: Expanded view of the round 2 Pareto front solutions including the Set0 solution (black dot) from the previous study of Deeth et al.²⁹

Once all of the new solutions, the children, have been created, we have doubled the population of solutions. The children are then used, and their fitness is assessed. The entire population of both parent and children solutions is then Pareto ranked and the density for each solution found. A new population is then made that has as many members as the initial population. The population members are chosen first in order of rank. If the next set of members for the next rank is too many, such that the new population would be too large, then the members of that rank with the lowest density are selected. It is this final new population of the strongest parent and children solutions that becomes the parents to the next set of solutions. And so, the process of mating, mutation, and selection is repeated for a number of generations.

The Pareto front of the final generation of solutions is a set of equally valid solutions that can be used in simulations. Ultimately, it is up to the user to choose which parameter set to use, their choice being guided by the requirements of the problem. So if the aim of the parametrization is to provide excellent structures, while energies are not as important, then solutions can be chosen that have lower structural errors while having higher energy errors compared to other solutions.

Rather than simply run a single training session for a large number of generations, we will be training the parameters progressively. Since we have elected that the parameters can only vary by a predefined amount during each training run, there is a limit to how much we can improve the parameters, and subsequently their performance. However, by using a member of the final generation Pareto front as the new origin for a new search, we can progressively move through the full parameter space, while exploring and optimizing parameters within a smaller region at each step. There is no predetermined number of generations for which a training session should be run. In general, the more parameters being trained, the larger the population of test sets, the longer it takes to converge on the Pareto front. Of course, it is inefficient to run a training session for more generations than is required. So, as a rule, the number of generations required for convergence has to be chosen by the user, on the basis of what they have already learnt about training the parameters beforehand. Equally, the larger the bit string used for the substrings, the longer it takes to explore parameter space and so converge. However, a fine grade search can then be performed. The same is true if the window of the parameter space search is

increased. Thus, all of these modifications have to be considered when determining the number of generations used.

RESULTS

Many transition metal centers can support more than one spin state. The most common examples are six-coordinate complexes of metal ions with formal d configurations spanning d^4 through to d^7 which generally display one of two magnetic states: high spin (HS), with the maximum number of unpaired electrons—4, 5, 4, and 3, respectively—or low spin (LS) with fewer unpaired electrons—2, 1, 0, and 1, respectively.

Under favorable circumstances, the HS and LS states may lie very close in energy such that an external influence, such as heat, pressure or light, causes the spin state to change. This so-called spin crossover (SCO) behavior confers a bistability which can potentially be exploited in the field of molecular electronics, data storage, and display devices.⁴²

The best known SCO complexes are d^6 Fe(II) systems.⁴³ Fe(II) complexes can display thermal SCO where the LS $^1A_{1g}$ state is favored at low temperature but the HS $^5T_{2g}$ state becomes favored at higher temperatures due to its greater entropy. Alternatively, some complexes can be excited from their ground state to an excited metastable state of different spin (so-called light induced excited spin state trapping or LIESST). The excited state can be stable indefinitely provided the temperature is kept low enough to prevent the thermal relaxation back to the ground state.

The spin state is determined by the balance between the $d-d$ interelectron repulsion and the ligand field stabilization energy (LFSE). The LFSE favors low spin; interelectron repulsion favors high spin. The latter is more or less constant for a particular metal ion, so the main variable that we can tune with the ligands is Δ_{oct} . Nitrogen donor ligands have the right ligand field strength with the first identified spin crossover system being $[\text{Fe}(\text{phen})_2(\text{NCE})_2]$ ($E = \text{S, Se}$).^{44–46}

The complexes shown in Figure 1 span the thermal spin crossover divide, with FEAM6 and PURYIK being HS in the solid state while DETTOL and PAZZAP are low spin. Using this data set, a force field was constructed, largely manually, which captured the spin state energetics as computed by DFT.²⁹ This FF was then used to design new complexes with small predicted

Table 1. Comparison of the DFT and LFMM Predicted Spin Crossover Energies ΔE (kcal mol⁻¹) Using the Previous LFMM Parameters (Set₀)²⁹ and Parameter Sets Discovered in the Second Round of Training Round of Parameter Searching^a

system	$\Delta E(\text{DFT})$	$\Delta E(\text{Set}_0)$	$\Delta E(\text{R2S4})$
FEAM6	6.7	4.71	4.89
PURYIK	1.6	1.99	2.20
DETTOL	-1.5	0.56	0.44
PAZXAP	-14.4	-13.97	-15.15
RMSD		<i>1.46</i>	<i>1.41</i>
MUE		<i>1.22</i>	<i>1.28</i>

^a Root mean square deviations (RMSDs) and the mean unsigned errors (MUEs) are included in italics.

spin state energy differences which could potentially display SCO behavior.

In the present work, the parameters are (re)optimized with respect to minimizing two objective functions—the root-mean-squared deviation (RMSD) in DFT versus LFMM spin state energy differences (kcal mol⁻¹), and the RMSD between the DFT and LFMM Fe–N and N–N distances (Å).

Our first test is to determine whether the new parametrization method can recover the existing parameter set, Set₀. Starting with the Morse function parameters—the dissociation energy, D , the reference metal–ligand distance, r_0 , and the curvature parameter, α —we significantly changed the Set₀ values from 58.3 kcal/mol, 2.15 Å, and 1.318 to 108.3 kcal/mol, 3.15 Å, and 2.10, respectively, to provide a poor starting guess (red dot in Figure 2).

In round 1 of optimization, an initial population of parameter sets is generated which conforms to the variation limits described above, and their performance is evaluated. They are then progressively refined over 50 generations to give a Pareto front of solutions, one of which was selected (the green triangle in Figure 2) to initiate round 2 of optimization. Thus, we shift the window of parameter space to allow for a progressive movement from a poor region of parameter space to a better one. We now reduce the allowed amount of variation such that the Morse functions reference distances, r_0 , are varied by up to 0.5 Å in either direction. The dissociation energies, D , of the functions are varied by up to 25% in either direction, while the α coefficients are varied by up to 25% in either direction. After a further 50 generations, we arrive at a new Pareto front of solutions. Table 1 compares spin state energy differences, $\Delta E = E_{LS} - E_{HS}$ from DFT (column 2) with the previously reported LFMM parameters (Set₀) and the fourth set from round 2 (R2S4). The latter has very similar RMSDs to Set₀, demonstrating that the automated MOEA procedure can quickly recover a good solution even when the optimization is started from a poor initial guess.

Table 2 compares the parameter sets we have discovered to the original parameters. The performance of the new parameters is as good as Set₀ (a more detailed comparison of the structural features is provided in the Supporting Information, Tables S2–S6); we would ideally prefer to have energy errors, on average, of less than a 1 kcal mol⁻¹. However, we find that further training using this new starting point does not lead to much further improvement. This is not surprising as this is simply the limit of performance that can be obtained varying just the Morse potential parameters.

Table 2. Comparison of the Parameter Set Values from the First Run of Parameterization and the Original Parameters, Set₀

parameter set	$D_c/\text{kcal mol}^{-1}$	$r_0/\text{Å}$	$\alpha/\text{Å}^{-1}$
Set ₀	58.3	2.15	1.318
R2S4	143.8	2.15	0.828

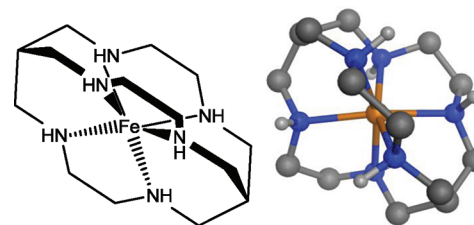


Figure 3. Structural representations of FE399.

Table 3. Comparison of the DFT and LFMM Predicted Spin State Energy Differences, Using the Original LFMM Parameters, Set₀, and Parameters Sets Obtained from Rounds 3, 4, and 5 of Optimization

system	DFT	Set ₀ ^a	R3S4 ^a	R4S3 ^b	R5S7 ^b
FEAM6	6.7	4.7	3.3	4.7	6.3
PURYIK	1.6	1.4	-0.2	1.8	1.8
DETTOL	-1.5	-0.5	-2.9	-0.4	-1.4
PAZXAP	-14.4	-16.0	-18.8	-16.3	-14.5
FE399	-1.6	12.2	9.8	1.4	-1.4
RMSD		6.3	5.8	1.9	0.2
MUE		3.7	4.5	1.6	0.2

^a Force field with original Fe–N–C–C torsion term. ^b Force field with modified Fe–N–C–C torsion.

We now apply the parameters to other iron amine complexes such as the cage species in Figure 3 (FE399), which is reported to undergo a spin transition in CD₃CN solution.⁴⁷

However, the R2S4 parameter set (and indeed Set₀) predicts that FE399 should be strongly high spin. In contrast, DFT suggests that FE399 should be low spin by 1.6 kcal mol⁻¹ and is thus consistent with potential SCO behavior. Compared to DFT, to which the LFMM parameters were tuned, the current set gives a massive energy error of nearly 14 kcal mol⁻¹. The implication is that some feature or features of FE399 are not properly covered by the systems in the training set. Consequently, we added FE399 to the original Set₀ training set and attempted to reoptimize the LFMM parameters.

The large initial error for FE399 leads to a large RMSD energy error for the whole set of 6.3 kcal mol⁻¹, but the geometrical error for FE399 is tiny. In round 3 of retraining, only the Morse function parameters were considered, but no appreciable improvement was observed (Table 3, R3S4). Extending the parameters which were varied to include ligand–ligand repulsion gave a significant improvement (Table 3, R4S3), and although the qualitative prediction for FE399 is incorrect, the DFT and LFMM ΔE values agree to within 3 kcal mol⁻¹.

We have shown that the training method can incorporate new information with little effort, and a force field can be retrained to

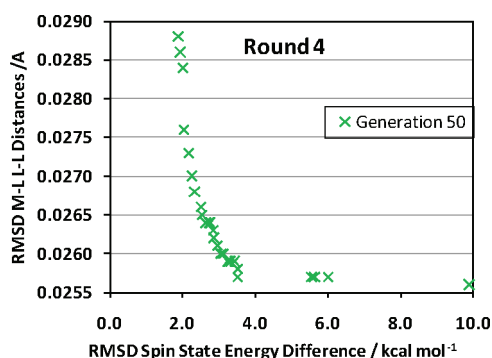


Figure 4. Final Pareto front solution from round 4 training using the extended database including FE399 but the original MMFF94 parameters.

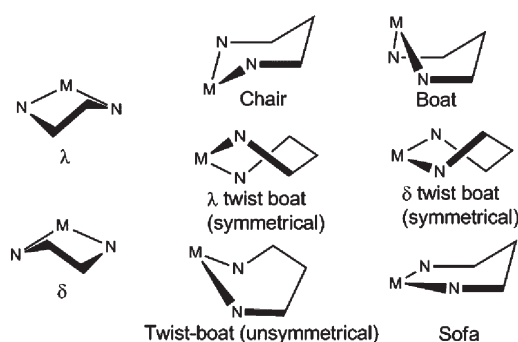


Figure 5. Conformations of five- and six-membered chelate rings.

give acceptable errors. However, we were puzzled by the observation of virtually perfect agreement between the LFMM and DFT HS and LS structures of FE399 while the ΔE error was relatively large. The Pareto front of round 4 (Figure 4) suggests a limiting RMSD error in ΔE of around 2 kcal mol⁻¹. Even if we allow for *all* of the LFMM parameters to be altered, we are unable to improve much further. This suggests that there may be a problem not with the LFMM parameters but with the MMFF94 force field. It turns out that FE399 exposes an issue with the Fe–N–C–C torsion term.

The structure of the cage ligand creates two sets of six-membered chelate rings at the top and bottom of the complex connected by five-membered rings in the middle. Five-membered chelates are relatively straightforward since there are only two possibilities, δ and λ (Figure 5, left), and their torsion angles do not vary much as the metal and/or bite angle varies. In contrast, the torsion angles around six-membered rings vary dramatically with conformation.

While there are many examples of LFMM applications to both five- and six-membered chelates, FE399 possesses two unusual features. First, the six-membered rings are all obliged to adopt a boat conformation, and second, this conformation is affected by the spin state. In HS, the boat is rather twisted (the Fe–N–C–C torsion angle, τ_{boat} is $\sim 29^\circ$), but the stronger ligand field for LS forces a near-perfect boat conformation with τ_{boat} equal to $\sim 4^\circ$ (Figure 6). Thus, whereas in other complexes, the HS/LS conformations are similar and any error arising from nonideal torsional parameters cancels out, FE399 specifically highlights any shortcomings in dealing with the chair/boat conformational

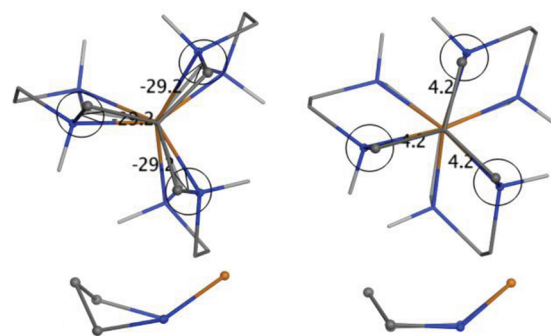


Figure 6. Detail showing the change in six-membered ring conformations going from high spin (left) to low spin (right). Nonpolar hydrogens omitted for clarity. Structures are DFT-optimized geometries.

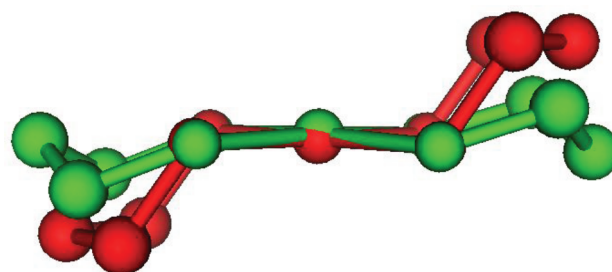


Figure 7. Calculated structure (red) of $[\text{Cu}(\text{pn})_2]^{2+}$ without torsional constraints compared to the experimental geometry (green) from ref 48. Hydrogen atoms have been removed for clarity.

energy difference. Thus, our usual approach of taking a generic “reference” $^*-\text{N}-\text{C}-^*$ torsional term (* represents any atom type) to describe the Fe–N–C–C torsion is found wanting.

By default, MOE assigns $\cos 2\tau$ and $\cos 3\tau$ terms for the Fe–N–C–C torsion which were satisfactory in the previous study but, for FE399, lead to a LS energy about 14 kcal mol⁻¹ too high. We can lower the LS energy by modifying the Fe–N–C–C torsional term. By adding a $\cos 4\tau$ term with a suitable force constant, the energy difference between $\tau = 30^\circ$ and $\tau = 0^\circ$ was reduced by about 0.5 kcal mol⁻¹ (See Figure S1, Supporting Information). Given that there are six FeN₂C₃ chelate rings in FE399, this has the required effect of reducing the LS energy by the necessary amount. However, this is at the expense of a reduction in the trigonal twist of the HS complex. Of course, as per our standard procedures, the N–Fe–N–C torsional potential has zero force constants since in octahedral complexes, *trans* nitrogen donors with a bond angle of 180° would result in an undefined τ . This is not a problem for five-membered chelates, but we have already seen its effect for six-membered rings in $[\text{Cu}(\text{pn})_2]^{2+}$ where the experimental conformation is much closer to sofa while the default MOE parameters yield a much more chairlike geometry⁴⁸ (Figure 7). This was remedied by adding explicit torsional restraints to help flatten out the pn ring. The same could have been done here except that the cage ligand constrains the rings already, plus the Fe–N–C–C torsion would need to be refined again, which seems unnecessary.

With the addition of the modified torsion term and subsequent retraining, we are able to improve the fit. In round 5, only the Morse function parameters were included, while in round 6, all LFMM parameters were reoptimized, but this did not improve

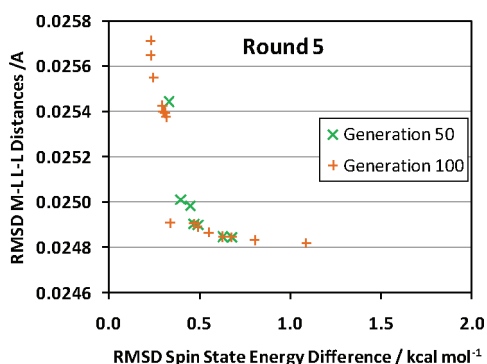


Figure 8. Pareto front from round 5 of optimization.

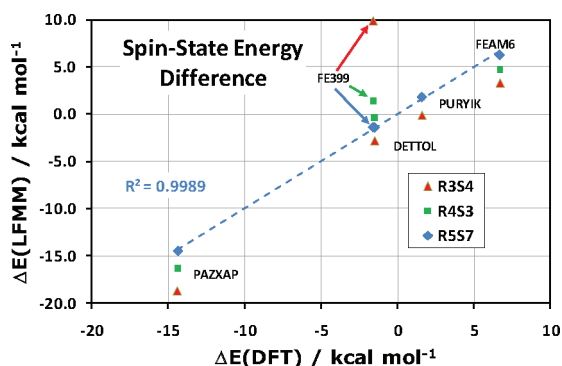


Figure 9. LFMM and DFT spin state energy difference for different parametrization scenarios: R3S4 excludes FE399 from training. R4S3 includes FE399 but retains the original MMFF94 Fe–N–C–C torsion term. R5S7 includes FE399 and the modified Fe–N–C–C torsion term.

the round 5 solutions. The final round 5 Pareto front solution is shown in Figure 8. The geometrical objective is confined to a narrow range of ~ 0.001 Å, while the energy objective varies by ~ 1 kcal mol $^{-1}$. The ΔE values for the R5S7 set are shown in Table 3.

Figure 9 compares the spin state energy differences for all five complexes (Figure 1 and FE399) for three scenarios: R3S5 excludes FE399 from training. R4S3 includes FE399 but retains the original MMFF94 Fe–N–C–C torsion term. R5S7 includes FE399 and the modified Fe–N–C–C torsion term. The latter shows that virtually perfect agreement between DFT and LFMM can be obtained using the MOEA method with the Pearson R^2 value for R5S7 of 0.9989.

Table 4 compares the LFMM parameter values for a selection of comparable round 5 solutions with the original Set $_0$. They are all broadly similar, although the MOEA sets tend to have smaller dissociation energy values and a lower Morse curvature. The a_5 term for the spin-pairing energy is also more negative.

Introducing the torsion term modification to the MMFF94 force field has had an influence on the Morse potential form, while having little effect on the ligand–ligand repulsion term, A_{LL} , and AOM parameters, e_o and e_{ds} . What is clear from Tables 4 and 2 is that we require more data for training of the parameters, in particular for simultaneous fitting of D and α . The force constant that characterizes the Morse function is proportional to Da^2 . This means that for all of our parameter sets, these two

Table 4. Comparison of the LFMM Parameters from the Final Pareto Front Where All Parameters Are Being Varied

parameter	R5S7	R5S1	R5S14	Set $_0$
$r_0/\text{Å}$	2.186	2.190	2.191	2.15
$D/\text{kcal mol}^{-1}$	47.8	56.0	55.1	58.3
α	1.126	1.027	1.026	1.318
$A_{LL}/\text{kcal mol}^{-1} \text{Å}^{-6}$	3916	3741	3735	3935
$e_o/\text{cm}^{-1} \text{Å}^{-5}$	412808	412566	413129	413000
$e_{ds}/\text{cm}^{-1} \text{Å}^{-6}$	125765	126125	126248	126030
$a_0(\text{pair})/\text{kcal mol}^{-1}$	14.3	14.5	14.5	14.5
$a_5(\text{pair})/\text{kcal mol}^{-1} \text{Å}^{-5}$	−55.5	−54.5	−55.7	−44
energy error RMSD	0.233	0.627	1.088	6.288

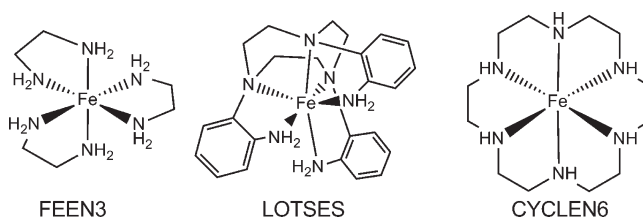


Figure 10. Structural diagrams of additional Fe(II) amine complexes.

Table 5. Calculated Spin State Energy Difference (kcal mol $^{-1}$) for Complexes in Figure 10

complex	$\Delta E(\text{DFT})$	$\Delta E(\text{LFMM})^a$	$\Delta \Delta E$
FEEN3	4.7	7.0	+2.3
LOTSES	3.2	2.8	−0.4
CYCLEN6	7.2	14.7	+7.5

^a R5S7 parameter set.

parameters can freely vary so long as this force constant remains the same, i.e., providing $Da^2 \sim 60$.

Having developed an extremely accurate LFMM FF for the training set, an obvious test is to use it to model iron amine complexes which were not used in any training. Three examples were chosen—[Fe(en) $_3$] $^{2+}$ (FEEN3, en = ethylenediamine), [Fe(1,4,7-tris(2-aminophenyl)-1,4,7-triazacyclononane)] $^{2+}$ (LOTSES), and [Fe(1,4,7,10,13,16-hexaazacyclooctadecane)] $^{2+}$ (CYCLEN6; Figure 10).

The calculated spin-state energy differences are collected in Table 5. While both theoretical methods predict that all three complexes should have high spin ground states and the detailed agreement between DFT and LFMM is satisfactory for FEEN3 and LOTSES, LFMM gives a much larger error for CYCLEN6. This correlates with the local structure around the metal center. The FeN $_6$ core of CYCLEN6 has the greatest deviation from octahedral symmetry, and thus the complexes used in training were not “diverse” enough to cope. This example highlights the issue of how best to choose the training data. We could include all possible systems in the training set which would give a good force field but would rapidly become unwieldy. A better solution is to develop a method which can automatically select from all possible training systems the “best”—i.e., most diverse—subset which will deliver the most accurate parameters most efficiently. This will be the subject of our next publication.

Meanwhile, we return to the systems in Figure 10 and compare the calculated results to experimental results. FEEN3 is straightforward and is reported to be a high spin complex in agreement with theory. However, for LOTSES, the single crystal X-ray diffraction structure is reported to have mean Fe–N distances of 2.10 Å and to be “predominantly low spin”, the latter based on EPR and magnetometry measurements. It is difficult to reconcile the observed bond lengths with a low-spin state where distances closer to 2 Å are expected and, indeed, calculated by DFT and LFMM. Unfortunately, no details of the magnetometry experiments are provided, but we speculate that a more consistent explanation of the experimental data is that LOTSES displays a spin equilibrium in the solid state with an approximately 50/50 mix of HS and LS at room temperature. Further experimental studies could help resolve this issue.

Finally, CYCLEN6 is reported to have a temperature-dependent effective magnetic moment which has been interpreted as due to “a reversible intramolecular electron transfer”⁴⁹ between an intermediate spin ($S = 1$) Fe(II) species and a high spin Fe(III)–radical anion species, which becomes increasingly important as the temperature rises above 130 K. Calculations do not support this interpretation. DFT optimization of the intermediate spin (IS) state, starting from crystallographic coordinates, gives a Jahn–Teller compressed structure some 15.8 kcal mol^{−1} above the ground HS state and 8.4 kcal mol^{−1} above the LS state. While there are no doubt other possible conformations, it does not seem likely that the IS could ever become the ground state. A more plausible explanation is that the materials are not pure Fe(II) complexes. Indeed, LOTSES is also reported to have a Fe(III) impurity of ~5%. Generating well-defined Fe(II) complexes free from paramagnetic impurities is clearly experimentally challenging.

CONCLUSIONS

We have shown that MOEAs are capable of fitting the LFMM parameters to reproduce DFT data to high accuracy with minimal energy and geometrical errors, despite starting from a parameter set which is a relatively poor guess. RMSD energy errors are reduced to just 0.2 kcal/mol with a geometrical RMSD error of 0.026 Å. Both objectives are significantly improved in a modest time (R1 training took 1 h for 30 generations, R5 took 5 h for 50 generations on a 64 bit 3 Ghz Intel Core 2 Duo laptop).

Minimizing one objective is in competition with the minimizing of the other objective. Minimizing both is ambiguous by traditional single objective methods, such as least-squares minimization, since a single parameter set is found, but this parameter set is in fact one of many, which are all equally valid and minimize the weighted sum of all errors.

MOEAs allow us to consider the payoff of optimizing one objective over another and to discover many different parameter sets which all minimize each of the objectives. The final selection of parameter sets which solve this problem is referred to as the Pareto front. All members of the Pareto front are no better than any of the other members of the Pareto front. For each pair of Pareto front solutions, both solutions will outperform the other for at least one objective. This is representative of there being no improvement in an objective without a loss of performance in another objective.

Using an implementation of MOEAs, the NSGA-II algorithm, we have not only been able to show that the method is able to recover parameters comparable to those found previously by

more laborious means but have also shown that the method is able to improve substantially upon these parameters, reducing energy errors by 75% and geometry errors by 90%. Furthermore, the method was used to reparameterize the LFMM force field with respect to an expanded set of complexes, highlighting the use of this tool to easily and iteratively improve and reparameterize a force field as more and more training data become available. Most importantly, this method can allow totally new force field parameters to be discovered for systems that have never been simulated before (this being a common occurrence within LFMM) and removes this tedious and time-consuming job from the user. However, it must be remembered that this method does not overcome inherent limitations of a force field, such as missing interaction potentials, and a user is still required to recognize this need and appropriately modify the force field. A user is also required to recognize when particular parameters do not play a role in improving the objectives and so direct the training method so that it may be more efficient. In terms of effort, it can now take just a few hours to reparameterize a force field, using a modest computer where only a handful of parameters are being altered. For larger problems, it only takes a couple of days, thus allowing new potentials to be parametrized rapidly.

In summary, we have shown, using Fe–amine complexes, that MOEAs are capable and efficient machine learning methods that can parametrize force fields as well as, if not better than, a person can do manually, or by single objective methods. They are also able, unlike traditional single objective training methods, to provide a number of possible solutions, from which the user can choose the most suitable, based upon their desired performance.

ASSOCIATED CONTENT

S Supporting Information. Figure S1: MMFF94 Fe–N–C–C torsion energy profiles for Fe–NH₂CH₂CH₃ moiety. Table S1: DFT-optimized Cartesian coordinates (Å) and ADF binding energies (kcal mol^{−1}) for complexes displayed in Figures 1, 3, and 10. Tables S2–S6: Fe–N and N–N distance comparisons for DFT and LFMM structures of FEAM6, PURYIK, DETTOL, PAZZAP (Figure 1), and FE399 (Figure 3) using the initial LFMM parameter set from ref 29. This information is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: r.j.deeth@warwick.ac.uk

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

The authors acknowledge the financial support of the EPSRC for the provision of a fellowship for CMH (Grant: EP/F042159) and access to the Chemical Database Service.⁵⁰

REFERENCES


(1) van Gunsteren, W. F.; Berendsen, H. J. C. *Groningen Molecular Simulation (GROMOS)*; University of Groningen: Groningen, The Netherlands, 1987.

- (2) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187.
- (3) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Profetajr, S.; Wiener, P. *J. Am. Chem. Soc.* **1984**, *106*, 765.
- (4) Jorgensen, W. L.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1988**, *110*, 1657.
- (5) Deeth, R. J.; Foulis, D. L. *Phys. Chem. Chem. Phys.* **2002**, *4*, 4292.
- (6) Deeth, R. J.; Anastasi, A.; Diedrich, C.; Randell, K. *Coord. Chem. Rev.* **2009**, *253*, 795.
- (7) Comba, P.; Remenyi, R. *Coord. Chem. Rev.* **2003**, *238–239*, 9.
- (8) Wang, J.; Kollman, P. A. *J. Comput. Chem.* **2001**, *22*, 1219.
- (9) Bowen, J. P.; Allinger, N. In *Rev. Comput. Chem.*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 1990; Vol. 9, p 81.
- (10) Hülsmana, M.; Müllerb, T. J.; Köddermana, T.; Reitha, D. *Mol. Simul.* **2010**, *36*, 1182.
- (11) Norrby, P. O.; Liljefors, T. *J. Comput. Chem.* **1998**, *19*, 1146.
- (12) Maple, J. R.; Hwang, M. J.; Stockfish, T. P.; Dinur, U.; Waldman, M.; Ewig, C. S.; Halger, A. T. *J. Comput. Chem.* **1993**, *15*, 162.
- (13) Warshel, A.; Lifson, S. *J. Chem. Phys.* **1970**, *53*, 582.
- (14) Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 490.
- (15) Allinger, N. L.; Yuh, Y. H.; Li, J.-H. *J. Am. Chem. Soc.* **1989**, *111*, 8551.
- (16) Norrby, P.-O.; Brandt, P. *Coord. Chem. Rev.* **2001**, *212*, 79.
- (17) Brandt, P.; Norrby, T.; Akermark, B.; Norrby, P.-O. *Inorg. Chem.* **1998**, *37*, 4120.
- (18) Marques, H. M.; Cukrowski, I. *Phys. Chem. Chem. Phys.* **2002**, *4*, 5878.
- (19) Marques, H. M.; Cukrowski, I. *Phys. Chem. Chem. Phys.* **2003**, *5*, 5499.
- (20) Skopec, C. E.; Robinson, J. M.; Cukrowski, I.; Marques, H. M. *J. Mol. Struct.* **2005**, *738*, 67.
- (21) De Sousa, A. S.; Fernandes, M. A.; Nxumalo, W.; Balderson, J. L.; Jeffic, T.; Cukrowski, I.; Marques, H. M. *J. Mol. Struct.* **2008**, *872*, 47.
- (22) Slepoy, A.; Peters, M. D.; Thompson, A. P. *J. Comput. Chem.* **2007**, *28*, 2465.
- (23) Mostaghim, S.; Hoffmann, M.; König, P. H.; Frauenheim, T.; Teich, J. In *IEEE Congress on Evolutionary Computation (CEC 2004)*; IEEE: Portland, OR, 2004; p 212.
- (24) Hunger, J.; Beyreuther, S.; Huttner, G.; Allinger, K.; Radelof, U.; Zsolnai, L. *Eur. J. Inorg. Chem.* **1998**, 693.
- (25) Hunger, J.; Huttner, G. *J. Comput. Chem.* **1999**, *20*, 455.
- (26) Tafipolsky, M.; Schmid, R. *J. Phys. Chem. B* **2009**, *113*, 1341.
- (27) Cundari, T. R.; Fu, W. *Inorg. Chim. Acta* **2000**, *300–302*, 113.
- (28) Nicolaou, C. A.; Brown, N.; Pattichis, C. S. *Curr. Opin. Drug Discovery Dev.* **2007**, *10*, 316.
- (29) Deeth, R. J.; Anastasi, A. E.; Wilcockson, M. J. *J. Am. Chem. Soc.* **2010**, *132*, 6876.
- (30) Zitzler, E.; Thiele, L. *IEEE Trans. Evol. Comp.* **1999**, *3*, 257.
- (31) Hawe, G.; Sykulski, J. *COMPEL* **2008**, *27*, 836.
- (32) Deb, K. In *Multiobjective Optimization*; Branke, J., Ed.; Springer-Verlag: Berlin, 2008; p 59.
- (33) Burnham, C. J.; Xantheas, S. S. *J. Chem. Phys.* **2002**, *116*, 1500.
- (34) Handley, C. M.; Hawe, G. I.; Kell, D. B.; Popelier, P. L. A. *Phys. Chem. Chem. Phys.* **2009**.
- (35) Deb, K.; Pratrap, A.; Agrawal, S.; Meyarivan, T. *IEEE Trans. Evol. Comp.* **2002**, *6*, 182.
- (36) Baerends, E. J.; Bérces, A.; Bo, C.; Boerrigter, P. M.; Cavallo, L.; Deng, L.; Dickson, R. M.; Ellis, D. E.; Fan, L.; Fischer, T. H.; Fonseca Guerra, C.; van Gisbergen, S. J. A.; Groeneveld, J. A.; Gritsenko, O. V.; Harris, F. E.; van den Hoek, P.; Jacobsen, H.; van Kessel, G.; Kootstra, F.; van Lenthe, E.; Osinga, V. P.; Philipsen, P. H. T.; Post, D.; Pye, C. C.; Ravenek, W.; Ros, P.; Schipper, P. R. T.; Schreckenbach, G.; Snijders, J. G.; Sola, M.; Swerhone, D.; te Velde, G.; Vernooijs, P.; Versluis, L.; Visser, O.; van Wezenbeek, E.; Wiesenekker, G.; Wolff, S. K.; Woo, T. K.; Ziegler, T. *ADF 2008.01*; Scientific Computing and Modelling NV, Free University, Amsterdam: Amsterdam, 2008.
- (37) Swart, M. *J. Chem. Theory Comput.* **2008**, *4*, 2057.
- (38) Hocking, R. K.; Deeth, R. J.; Hambley, T. W. *Inorg. Chem.* **2007**, *46*, 8238.
- (39) Merbach, A. E. *Pure Appl. Chem.* **1987**, *59*, 161.
- (40) Deeth, R. J.; Fey, N.; Williams-Hubbard, B. J. *J. Comput. Chem.* **2005**, *26*, 123.
- (41) MOE Molecular Operating Environment, 2010.10; Chemical Computing Group: Montreal, Canada, 2010.
- (42) Letard, J. F.; Guionneau, P.; Goux-Capes, L. In *Spin Crossover in Transition Metal Compounds III*; Guetlich, P., Goodwin, H. A., Eds.; Springer: 2004; Topics in Current Chemistry Vol. 235, p 221.
- (43) Halcrow, M. A. *Polyhedron* **2007**, *26*, 3523.
- (44) König, E.; Madeja, K. *Chem. Commun.* **1966**, 61.
- (45) König, E.; Madeja, K. *Inorg. Chem.* **1967**, *6*, 48.
- (46) Baker, W. A.; Bobonich, H. M. *Inorg. Chem.* **1964**, *3*, 1184.
- (47) Martin, L. L.; Hagen, K. S.; Hauser, A.; Martin, R. L.; Sargeson, A. M. *J. Chem. Soc., Chem. Commun.* **1988**, 1313.
- (48) Deeth, R. J.; Hearnshaw, L. J. A. *Dalton Trans.* **2005**, 3638.
- (49) Mitewa, M.; Bontchev, P. R.; Russanov, V.; Zhecheva, E.; Mechandjiev, D.; Kabassanov, K. *Polyhedron* **1991**, *10*, 763.
- (50) Fletcher, D. A.; McMeeking, R. F.; Parkin, D. J. *Chem. Inf. Comput. Sci.* **1996**, *36*, 746.

RASPT2/RASSCF vs Range-Separated/Hybrid DFT Methods: Assessing the Excited States of a Ru(II)bipyridyl Complex

Daniel Escudero[†] and Leticia González^{*,‡}

Institut für Physikalische Chemie, Friedrich-Schiller Universität, Helmholtzweg, 4, 07743 Jena, Germany

 Supporting Information

ABSTRACT: The excited states of the *trans*(Cl)-Ru(bpy)Cl₂(CO)₂ (bpy = bipyridyl) transition-metal (TM) complex are assessed using the newly developed second-order perturbation theory restricted active space (RASPT2/RASSCF) method. The delicate problem of partitioning the RAS subspaces (RAS1, RAS2, and RAS3) is addressed, being the choice of the RAS2 the bottleneck to obtain a balanced description of the excited states of different nature when TMs are present. We find that the RAS2 should be composed by the correlation orbitals involved in covalent metal–ligand bonds. The level of excitations within the RAS1 and RAS3 subspaces is also examined. The performance of different flavors of time-dependent density functional theory including pure, hybrid, meta-hybrid, and range-separated functionals in the presence of solvent effects is also evaluated. It is found that none of the functionals can optimally describe all the excited states simultaneously. However, the hybrid M06, B3LYP, and PBE0 functionals seem to be the best compromise to obtain a balanced description of the excited states of *trans*(Cl)-Ru(bpy)Cl₂(CO)₂, when comparing with the experimental spectrum. The conclusions obtained in this molecule should pave the road to properly treat excited states of larger Ru–polypyridyl complexes, which are of particular interest in supramolecular chemistry.

1. INTRODUCTION

Ru(II) polypyridyl complexes and related compounds are promising candidates as light-harvesting antennas in hot research areas, such as artificial photosynthesis,¹ light-driven catalysis (i.e., sunlight driven splitting of water),² or dye-sensitized solar cells (DSSCs),³ due to a combination of optimal chemical, electrochemical, and photophysical properties. The identification and characterization of the lowest-lying excited states are therefore a challenging task of paramount importance to guide the design of molecular functional materials. To get an insight into the photophysical properties of these complexes, the use of accurate *ab initio* multiconfiguration methods, as for example, the well-established second-order perturbation theory complete active space⁴ (CASPT2/CASSCF) protocol, is highly desirable. Due to the extensive size of such transition-metal (TM) complexes and the need of large active spaces able to handle all the static correlation, such calculations are yet pretty much at the limit of the current computational resources and not very much extended.⁵ To our knowledge, there are only a few examples of CASPT2/CASSCF studies on TM–polypyridyl complexes, e.g., on [Fe(bpy)₃]²⁺⁶ or [Re(bpy)(*t*-stpy)]⁺.⁷ Despite the considerable effort involved in these studies, they involve reduced CAS reference wave functions with active spaces that might not be extensive enough to account for all the desired correlation effects. The restricted active space method (RASSCF)⁸ and its PT2 extension (RASPT2)⁹ are very appealing because they allow using considerably larger active spaces than the CASPT2/CASSCF protocol. However, due to the three different partitions of the active space, a number of open questions can be raised regarding its systematic use. In particular, it is not straightforward how the active orbitals in the RAS subspaces should be best distributed or at which level of excitation the results can be considered converged. Recently, Gagliardi and co-workers have performed an extensive number of RASPT2/RASSCF calculations mainly in

organic dyes,^{10,11} clearly illustrating that the selection of the RAS spaces requires careful calibration. In the case of simple oligomeric π -conjugated systems, the computationally cheapest strategy is leaving the RAS2 empty while allowing for single, double, triple, and quadruple (SDTQ) excitations within the RAS1 and RAS3 subspaces. This simple recipe allows for an accurate description of ionization potentials and lowest-lying excited states of many organic systems,¹⁰ but unfortunately it cannot be extrapolated to more complicated systems, neither of organic character, like free base porphyrins,¹¹ nor of inorganic nature, like Cu(I)- α -ketocarboxylate complexes.¹² From this perspective, there is an urge to calibrate the RASSCF approach in TM complexes which otherwise are typically treated by means of density functional theory (DFT) and its time-dependent version (TD-DFT).¹³

The compromise between low computational cost and accuracy obtained with DFT is in general remarkable, and therefore a huge number of ground-state studies for TM complexes are found in the literature.¹⁴ The success of DFT describing ground-state properties is however not comparable to the success of TD-DFT for the study of excited states. TD-DFT is known to have difficulties in describing Rydberg states, charge-transfer (CT) excitations¹⁵ as well as doubly or highly excited states.¹⁶ Attempts to improve the description of CT and Rydberg excitations, while maintaining good quality for local excitations, has led to the development of hybrid functionals with intermediate percentage of exact exchange, such as PBE0,¹⁷ and range-separated hybrid functionals, such as LC- ω PBE¹⁸ or CAM-B3LYP,¹⁹ which have been shown to perform reasonably well describing CT of organic dyes.²⁰ CT events are common in TM spectroscopy; however, more challenging is that the calculation of the UV–vis spectra of

Received: September 13, 2011

Published: November 09, 2011

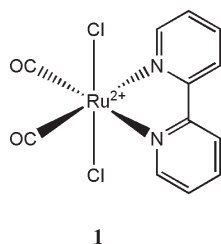


Figure 1. Chemical structure of *trans*-(Cl)-Ru(bpy)Cl₂(CO)₂, complex 1.

complex TM systems also requires the simultaneous balanced description of additional intraligand (IL) and d–d transitions, which are both local excitations. Despite these difficulties, TD-DFT is indiscriminately employed due to its simplicity and apparent black-box behavior, in particular for large systems such as Ru(II) polypyridyl complexes,²¹ which are otherwise out of reach from more accurate *ab initio* methods. Few examples are found in the literature assessing the performance of TD-DFT on TM complexes.²² However, to our knowledge, there is no evaluation of the performance of hybrid vs range-separated functionals to describe the excited states of TM–polypyridyl and related complexes.

In this contribution we present a RASPT2/RASSCF study of the electronically excited states of *trans*-(Cl)-Ru(bpy)Cl₂(CO)₂ for which experimental data are available for comparison.^{23,24} This complex serves as an example of this general class of systems. It displays different types of excited states and thus represents a challenge for computational chemistry. Different partition schemes of the RAS subspaces (RAS1, RAS2 and RAS3) as well as the level of excitations in the RAS1 and RAS3 subspaces are evaluated. The conclusions reached with this molecular system will provide hints about how to study similar TM compounds, therefore, extending the possibilities of RASPT2/RASSCF to larger TM–polypyridyl complexes. Additionally, and based on the experimental results, the performance of TD-DFT using different functionals in solution is analyzed. Particularly, the behavior of several hybrid, meta-hybrid, pure, and long-range corrected functionals in describing the different types of excitations present in TM complexes is examined and discussed. The conclusions should help to make an adequate choice of the functional when studying very large TM–polypyridyl complexes that cannot be treated within the RASPT2/RASSCF protocol.

2. COMPUTATIONAL DETAILS

The *trans*-(Cl)-Ru(bpy)Cl₂(CO)₂ complex **1** (see Figure 1) was optimized in its electronic ground state under the C_{2v} symmetry constraint at the B3LYP/6–31G* level of theory. Relativistic effects in the Ru atom were considered using the ECP-28-mwb pseudopotential.²⁵ The complex has been characterized as a true minimum by calculating the Hessian at the same level of theory.

Single point CASPT2/CASSCF and RASPT2/RASSCF calculations were performed on the C_{2v} geometry. These calculations were done with the ANO-rcc-VTZP basis set.²⁶ Scalar relativistic effects were considered using a standard second-order Douglas–Kroll–Hess (DKH) Hamiltonian.²⁷ Cholesky decomposition of the electron repulsion integral matrix²⁸ was used, then reducing the computational times and the disk storage needs. In the CAS calculations, the traditional notation is used, namely CAS(*n*,*j*), where *n* is the number of electrons included in the

active space and *j* is the number of active orbitals. In the RAS calculations, the RAS(*n*,*l*,*m*; *i*,*j*,*k*) notation is employed, where *n* is the number of active electrons, *l* the maximum of holes in the RAS1, *m* the maximum of electrons in RAS3, and *i*, *j*, and *k* the number of orbitals in RAS1, RAS2, and RAS3, respectively.

The choice of the active orbitals for the CASSCF/RASSCF calculations is made in terms of the standard rules for TM compounds.^{29,30} Important correlation effects due to the covalency of the Ru metal–ligand bonds via σ - and π -bonding interactions are considered by including some relevant σ orbitals involving the Ru atom and also the chlorine atoms ($n_{\text{Cl}}-a_1$ orbital) or the bpy and the CO ligands ($\sigma_{\text{CO-bpy}}$ orbitals, in Figure 2). Relevant π -bonding interactions are taken into account by including π^*_{CO} orbitals and lone pairs of the chlorines atoms ($n_{\text{Cl}}-b_1$ orbital). In principle we follow the advice given by Pierloot³¹ that all orbitals with an important contribution of d character should be included in the active space. Since Ru is a 4d atom, the “double-shell” effect is not as indispensable as for 3d atoms; therefore, this external shell is not considered here for the sake of computational saving. In some sense, these correlation effects are partially recovered upon inclusion of other orbitals, such as the π^*_{CO} orbitals, which possess some contribution of 5d character. The intershell effects, i.e., those due to semicore electrons, are very important for the 4d and 5d series, and thus they have been inherently considered by using the ANO-rcc basis set that is optimized to include such correlation effects. Other valence orbitals, such as the π_{CO} orbitals and some $\pi_{\text{bpy}}/\pi^*_{\text{bpy}}$ orbital pairs, are not included in our calculations since a full valence electron treatment is out of reach for CASPT2 and even for RASPT2 calculations in the complex **1**. One strategy followed here is to include many of the orbitals that in principle contribute to static correlation but to a lesser extent, in the RAS1 and RAS3 subspaces, while RAS2 includes the indispensable orbitals involved in the main excited states (low- and high-lying) of **1**; these are the five 4d orbitals and the $\pi_{\text{bpy}}/\pi^*_{\text{bpy}}$ orbitals. These orbitals are involved in the main metal-centered (MC), metal-to-ligand charge transfer (MLCT) and IL states of **1**. An additional strategy is to include the relevant orbitals involved in the main electronic excitations within the RAS1/3 subspaces, while the correlation orbitals, i.e., those participating in the covalency of the Ru metal–ligand bonds via σ - and π -bonding interactions, will be part of the RAS2 subspace, since the correlation orbitals are essential to recover static correlation on the zeroth-order wave function. All the relevant orbitals are depicted in Figure 2.

Specifically, the following subspaces were constructed:

- Small CAS active space. It is composed of the five 4d orbitals of the Ru atom and a balanced set of four frontier $\pi_{\text{bpy}}/\pi^*_{\text{bpy}}$ pair of orbitals, see Figure 2. This active space is used in the CASSCF(14,13) calculations.
- Inclusion of correlation effects to the small CAS active space. A relevant n_{Cl} orbital with important σ character, involving lone pairs of the chlorine and the Ru center (see $n_{\text{Cl}}-a_1$ in Figure 2) as well as two π^*_{CO} orbitals ($\pi^*_{\text{CO}}-b_1$ and $\pi^*_{\text{CO}}-a_2$) are included in the active space through the RAS1 and RAS3 subspaces. Such selection was made in terms of the higher metal d contributions of such orbitals. Additionally, a pair of $\pi_{\text{bpy}}/\pi^*_{\text{bpy}}$ orbitals (namely the $\pi_{\text{bpy}}-2a_2/\pi^*_{\text{bpy}}-2a_2$ orbitals, see Figure 2) is shifted to the RAS1 and RAS3 subspaces, respectively, because the corresponding occupation numbers were higher/lower than 1.95/0.05. This active space is used in the RASSCF(16,2,2;2,11,3),

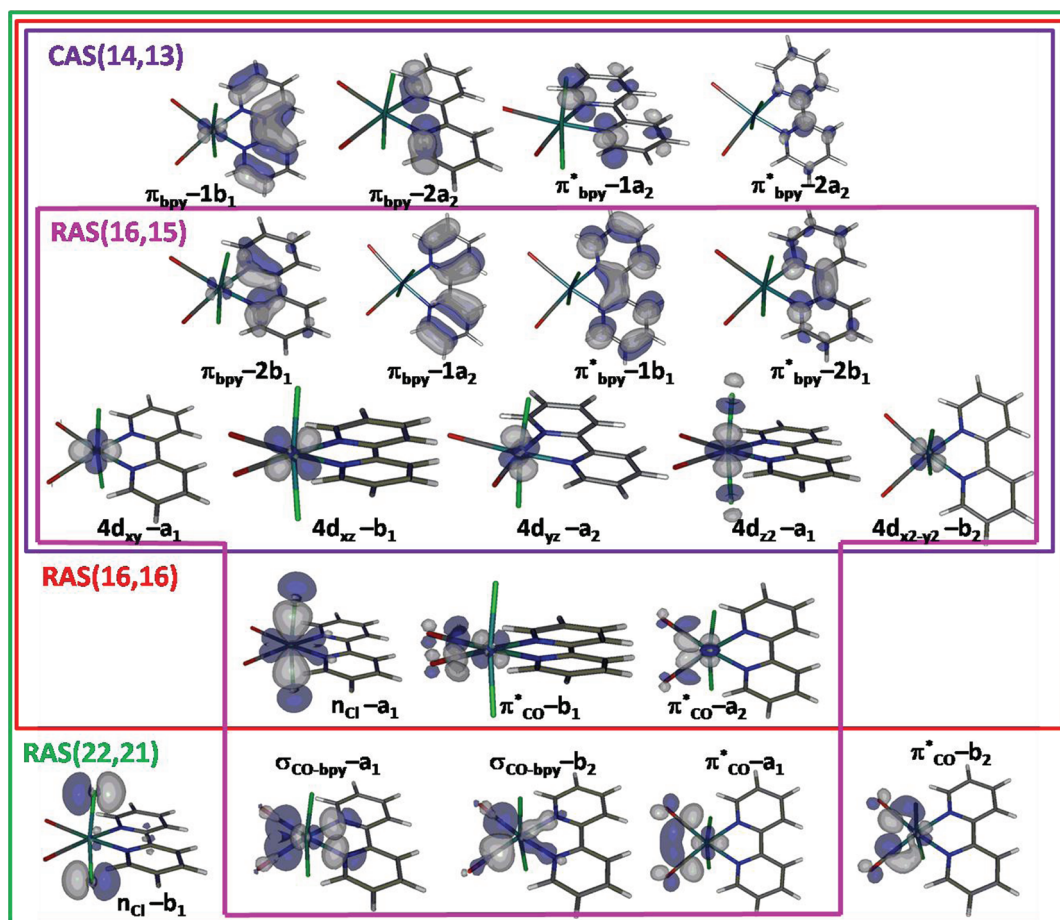


Figure 2. CASSCF and RASSCF active spaces employed.

RASSCF(16,3,3;2,11,3), and RASSCF(16,3,4;2,11,3) calculations. For comparison, a calculation where the RAS2 subspace is empty and all the active orbitals are assigned to the RAS1 and RAS3 subspaces is also performed, RASSCF(16,4,4;8,0,8). For the same reason as before, an additional pair of $\pi_{\text{bpy}}/\pi^*_{\text{bpy}}$ orbitals (namely the pair $\pi_{\text{bpy}}-1b_1/\pi^*_{\text{bpy}}-1a_2$) was moved into the RAS1/3 subspaces, leading to the RASSCF(16,2,2;3,9,4) calculation.

- (iii) Additional correlation effects: Based on the RAS(16,2,2;3,9,4) active space, additional correlation orbitals were considered. Again, such selection was done paying attention to the metal d contribution of the included orbitals: In the RAS1 an additional n_{Cl} orbital ($n_{\text{Cl}}-b_1$) and two $\sigma_{\text{CO-bpy}}$ orbitals ($\sigma_{\text{CO-bpy}}-a_1$ and $\sigma_{\text{CO-bpy}}-b_2$) were included, and the two remaining π^*_{CO} orbitals ($\pi^*_{\text{CO}}-a_1$ and $\pi^*_{\text{CO}}-b_2$) were put in the RAS3 (see Figure 2). In these calculations the two $\pi_{\text{bpy}}/\pi^*_{\text{bpy}}$ orbital pairs, $\pi_{\text{bpy}}-2a_2/\pi^*_{\text{bpy}}-2a_2$ and $\pi_{\text{bpy}}-1b_1/\pi^*_{\text{bpy}}-1a_2$, were maintained in the RAS1/3 subspaces because they do not play a role in the low and high-lying excited states. This partition leads to a RASSCF(22,2,2;6,9,6) calculation.
- (iv) $\sigma-\pi$ correlation effects in RAS2: Based on the assumption that correlation orbitals should be included in the RAS2 subspace, this subspace is composed by the n_{Cl} orbital ($n_{\text{Cl}}-b_1$), two $\sigma_{\text{CO-bpy}}$ orbitals ($\sigma_{\text{CO-bpy}}-a_1$ and

$\sigma_{\text{CO-bpy}}-b_2$) and three π^*_{CO} orbitals ($\pi^*_{\text{CO}}-b_1$, $\pi^*_{\text{CO}}-a_2$, and $\pi^*_{\text{CO}}-a_1$). Since the occupied a_1 orbitals were partially mixed among them, also the $4d_{xy}-a_1$ orbital was included into the RAS2. The rest of 4d orbitals and only the relevant $\pi_{\text{bpy}}/\pi^*_{\text{bpy}}$ orbital pairs, namely $\pi_{\text{bpy}}-2b_1/\pi_{\text{bpy}}-1a_2$ and $\pi^*_{\text{bpy}}-1b_1/\pi^*_{\text{bpy}}-2b_1$, were included in the RAS1 and RAS3 subspaces, allowing up to SDT excitations. This partition leads to a RASSCF(16,3,3;4,7,4) calculations.

The CASSCF/RASSCF calculations presented herein are done as state average (SA-CASSCF/RASSCF) with equal weights. The subsequent CASPT2 calculations are reported as single-state (SS-) and multistate (MS-CASPT2).³² The RASPT2 results are only reported in the MS fashion. The number of roots employed in each irreducible representation is specified in the corresponding table caption. Oscillator strengths were calculated at both the SA-CASSCF, MS-CASPT2, and MS-RASPT2 levels of theory through the RAS state interaction method.³³ The core shell orbitals were kept frozen in the CASPT2 calculations in order to avoid BSSE errors. In all our CASPT2/RASPT2 calculations, an ionization potential electron affinity (IPEA) shift (default value of 0.25 au) for the zeroth-order Hamiltonian was employed.³⁴ Additionally, an extra level shift³⁵ of 0.3 au was used to prevent weakly coupling intruder states interference, mainly with the high-lying states. Similar values of ca. 0.25 au have been used elsewhere to compute excited states of TM complexes.⁹ The spin-orbit (SO) UV-vis spectra have been also computed

Table 1. Relative SA-CASSCF(14,13) and SS- and MS-CASPT2(14,13) as well as SO-MS-CASPT2(14,13) Electronic Transitions Energies, ΔE (in eV), with Oscillator Strengths f , and Main Assignment for Complex **1**^a

state	SA-CASSCF(14,13) ^b			SS-CASPT2 (14,13)		MS-CASPT2(14,13) ^c		SO-MS-CASPT2(14,13) ^c	
	assignment/weight (c^2)	f	ΔE (eV)	ΔE (eV)	ΔE (eV)/ c^2	f^d	ΔE (eV)	f	
S ₀ (1 ¹ A ₁)	¹ GS/0.86			0.0	0.0/0.86		0.0		
S ₁ (1 ¹ A ₂)	¹ MC (4d _{yz} -a ₂ →4d _{z²} -a ₁)/0.71	0.000	3.15	3.76	3.76/0.72	0.000 (0.000)	3.78	0.000	
S ₂ (1 ¹ B ₁)	¹ MC (4d _{xz} -b ₁ →4d _{z²} -a ₁)/0.79	0.009	3.59	3.49	3.49/0.77	0.009 (0.007)	3.52	0.009	
S ₃ (2 ¹ B ₁)	¹ MC (4d _{yz} -a ₂ →4d _{x²-y²} -b ₂)/0.78	0.006	4.77	4.19	4.19/0.76	0.004 (0.004)	4.23	0.003	
S ₄ (2 ¹ A ₂)	¹ MC (4d _{xz} -b ₁ →4d _{x²-y²} -b ₂)/0.75	0.000	4.82	4.81	4.82/0.77	0.000 (0.000)	4.82	0.000	
S ₅ (2 ¹ A ₁)	¹ MC (4d _{xy} -a ₁ →4d _{z²} -a ₁)/0.84	0.007	5.13	5.57	5.57/0.85	0.006 (0.007)	5.59	0.006	
S ₆ (1 ¹ B ₂)	¹ MLCT (4d _{yz} -a ₂ → π^* _{bpy} -2b ₁)/0.76	0.000	5.19	3.81	3.73/0.74	0.000 (0.000)	3.74	0.000	
S ₇ (2 ¹ B ₂)	¹ MC (4d _{xy} -a ₁ →4d _{x²-y²} -b ₂)/0.82	0.001	5.32	5.49	5.55/0.62	0.070 (0.018)	5.57	0.068	
S ₈ (3 ¹ B ₂)	¹ IL (π _{bpy} -1a ₂ → π^* _{bpy} -1b ₁)/0.28	0.195	5.44	5.40	5.49/0.30	0.049 (-) ^e	5.51	0.051	
	¹ IL (π _{bpy} -1a ₂ → π^* _{bpy} -2b ₁)/0.22								
S ₉ (3 ¹ A ₁)	¹ MLCT (4d _{xz} -b ₁ → π^* _{bpy} -2b ₁)/0.85	0.045	5.61	3.35	3.35/0.85	0.025 (0.059)	3.36	0.025	
S ₁₀ (4 ¹ B ₂)	¹ MLCT (4d _{yz} -a ₂ → π^* _{bpy} -1b ₁)/0.69	0.162	6.34	4.74	4.75/0.60	0.199 (0.012)	4.77	0.198	
S ₁₁ (5 ¹ B ₂)	¹ IL (π _{bpy} -1a ₂ → π^* _{bpy} -2b ₁)/0.52	0.405	6.64	4.65	4.49/0.40	0.306 (0.309)	4.50	0.306	

^a The square of the configuration interaction coefficient, c^2 , indicates the weight in the wave function of the leading CSF obtained by the indicated electron replacement. ^b SA-(3,3,2,6)-CASSCF(14,13) calculations for A₁, B₁, A₂, and B₂ symmetries, respectively. ^c SO-MS/MS-(3,3,2,6)-CASPT2(14,13) calculations for A₁, B₁, A₂, and B₂ symmetries, respectively. ^d In parenthesis, oscillator strengths obtained with MS-RASPT2(16,2,2;3,9,4). ^e This state was not computed at the MS-RASPT2(16,2,2;3,9,4) level of theory.

with the CAS(14,13) wave function. The SO couplings have been evaluated with the SO-RASSI approach³⁶ (further details can be found in the Supporting Information).

The TD-DFT calculations were performed at the C_{2v} geometry with a 6-311G* basis set (ECP-28-mwb pseudopotential for Ru). The different functionals employed are: (i) several hybrid functionals with an increasing amount of exact exchange in the following order: B3LYP³⁷ (20% of exact exchange), PBE0¹⁷ (25%), and B3LYP-35³⁸ (35%); (ii) the meta-hybrid M06³⁹ and M06-2X³⁹ functionals (with 27% and 54% of exact exchange, respectively); (iii) the pure functional PBE;⁴⁰ and (iv) the long-range corrected CAM-B3LYP and LC- ω PBE functionals. Additionally, TD-DFT calculations were also performed in solution using CH₃CN as solvent with the polarization continuum model,⁴¹ i.e., PCM-TD-DFT calculations, with the same basis set.

The ground-state optimization and TD-DFT calculations have been performed with the Gaussian09⁴² program package, while CASSCF/CASPT2 and RASSCF/RASPT2 calculations have been performed with the MOLCAS7.2⁴³ software.

3. RESULTS AND DISCUSSION

First of all, we report here the experimental values known for the complex **1**. The UV-vis spectrum recorded in CH₃CN is characterized by a small band peaking at 352 nm or 3.52 eV (3.62 eV in CH₂Cl₂) showing an intensity of 1550 M⁻¹cm⁻¹ and by a higher broader band centered at ca. 300 nm or 4.14 eV, with a higher relative intensity of 14 100 M⁻¹cm⁻¹.^{23,24} The peaks of this main band with their associated intensities are the following: 3.96 eV (14 100 M⁻¹cm⁻¹), 4.13 eV (11 300 M⁻¹cm⁻¹), and 4.34 (10 000 M⁻¹cm⁻¹) Taking these values as a reference, hereafter we shall discuss the values obtained theoretically with the methods described above.

3.1. CASPT2/CASSCF Calculations. Table 1 shows the CASPT2/CASSCF(14,13) results for the excited states of complex **1**. The CASSCF(14,13) calculations indicate that the

low-lying singlet electronic excitations are mainly weakly absorbing ¹MC states of different symmetries, see S₁–S₅ and S₇ states at the CASSCF level of theory. We note that among them there are states of A₂ symmetry that, although strictly forbidden by symmetry, might be populated, e.g., in the course of photochemical deactivation channels. The ¹MLCT and ¹IL states showing substantial oscillator strengths are the S₈, S₉, S₁₀, and S₁₁ states at the CASSCF level of theory. The ¹MLCT states consist of transitions from the 4d orbitals of the Ru atom to the π^* _{bpy} orbitals. The ¹IL states are transitions within the π _{bpy}/ π^* _{bpy} orbitals, and thus excitations retain certain local character. To describe them all simultaneously at the CASSCF level of theory, we need to compute at least six states of B₂ symmetry and 3/3/2 states of A₁/B₁/A₂ symmetry, respectively. The chosen number of states corresponds to the minimum amount of roots necessary to describe the spectrum below ca. 5 eV in a balance manner, taking into account that some spectroscopic states do not appear as low-energy roots at the CASSCF level of theory and that the final states involved in these transitions have mixed character (see, e.g., S₈ and S₁₁ in Table 1). Table 1 also contains the SS-CASPT2 and MS-CASPT2 values. Similar energies are obtained with both procedures, being the roots of the B₂ irreducible representation the ones most affected by the MS procedure, due to the larger mixing present at the SA-CASSCF level of theory. The inclusion of dynamical correlation via CASPT2 leads to a large stabilization of some ¹MLCT and ¹IL states and thus to state reordering. As a consequence, the S₉ state (3¹A₁) at the CASSCF level of theory, for example, becomes the first excited state with MS-CASPT2 (or SS-CASPT2). On the contrary, dynamical correlation does not severely affect the relative energies of the ¹MC states, which remain almost unaffected (e.g., S₂ and S₄) or are slightly blue/red-shifted (S₁ and S₃, respectively). Returning our attention to the 3¹A₁ state, this is predicted at 3.35 eV at CASPT2 level of theory and hence in reasonable agreement with the weak band peaking experimentally at ca. 3.52 eV. Note that this state is the lowest at CASPT2 level of theory, while it was the S₉ with CASSCF.

Table 2. Selected TD-DFT Electronic Transitions Energies, ΔE (in eV), with Oscillator Strengths f against Experimental Values^a

state	meta-hybrid functionals (% HF exchange)				range-separated functionals				hybrid functionals (% HF exchange)						pure functionals		
	M06 (27%)		M06-2X (54%)		CAM-B3LYP		LC- ω PBE		B3LYP-35 (35%)		PBE0 (25%)		B3LYP (20%)		PBE		exptl
	ΔE	f	ΔE	f	ΔE	f	ΔE	f	ΔE	f	ΔE	f	ΔE	f	ΔE	f	
1^1B_1 (1^1MC)	2.99	0.001	3.06	0.003	3.39	0.003	3.50	0.003	3.34	0.003	3.31	0.002	3.25	0.001	3.09	0.002	
	<i>3.00</i>	<i>0.002</i>	<i>3.02</i>	<i>0.003</i>	<i>3.38</i>	<i>0.004</i>	<i>3.46</i>	<i>0.004</i>	<i>3.33</i>	<i>0.004</i>	<i>3.34</i>	<i>0.004</i>	<i>3.30</i>	<i>0.003</i>	<i>3.19</i>	<i>0.003</i>	
3^1A_1 (1^1MLCT)	2.48	0.011	3.64	0.009	3.43	0.014	4.19	0.022	3.10	0.010	2.60	0.010	2.41	0.009	1.56	0.008	3.52 ^b
	<i>3.47</i>	<i>0.023</i>	<i>4.55</i>	<i>0.019</i>	<i>4.41</i>	<i>0.031</i>	<i>5.05</i>	<i>0.047</i>	<i>4.03</i>	<i>0.021</i>	<i>3.51</i>	<i>0.019</i>	<i>3.42</i>	<i>0.019</i>	<i>2.50</i>	<i>0.016</i>	3.62 ^c
5^1B_2 (1^1IL)	4.29	0.142	4.76	0.246	4.67	0.258	4.82	0.289	4.56	0.304	4.44	0.199	4.29	0.064	4.02	0.064	
	<i>4.18</i>	<i>0.315</i>	<i>4.63</i>	<i>0.358</i>	<i>4.53</i>	<i>0.386</i>	<i>4.64</i>	<i>0.353</i>	<i>4.42</i>	<i>0.392</i>	<i>4.32</i>	<i>0.280</i>	<i>4.21</i>	<i>0.247</i>	<i>3.94</i>	<i>0.287</i>	3.96 ^b
4^1B_2 (1^1MLCT)	3.35	0.002	4.66	0.067	4.51	0.016	5.65	0.014	4.08	0.001	3.49	0.002	3.25	0.001	2.19	0.002	4.13–4.34 ^b
	<i>4.42</i>	<i>0.071</i>	<i>5.77</i>	<i>0.022</i>	<i>5.71</i>	<i>0.034</i>	<i>6.68</i>	<i>0.139</i>	<i>5.01</i>	<i>0.029</i>	<i>4.49</i>	<i>0.107</i>	<i>4.36</i>	<i>0.123</i>	<i>3.21</i>	<i>0.002</i>	

^a Numbers in italics are calculated in solution (CH₃CN) with the PCM model. ^b In CH₃CN, from ref 23. ^c In CH₂Cl₂, from ref 24.

This demonstrates the importance of dynamical correlation on the different type of states, as explained above. For simplicity, and although the order of states is altered in many cases with the addition of the PT2 correction, in the following discussion we keep the state numbers as provided by the CASSCF method. In view of the high oscillator strengths obtained with CASSCF for the S_{11} (5^1B_2) and S_{10} (4^1B_2) states, theoretically predicted at 4.49 and 4.75 eV at the MS-CASPT2 level of theory, these states can be then considered responsible of the strong band peaking at 3.96 and 4.13–4.34 eV. These states are ca. 0.5 eV blue shifted with respect to the experiment, but one should keep in mind that the experimental data are obtained in the presence of solvent, not included in the present calculations.

Regarding the intensities of the calculated states, as it can be seen in Table 1, all the 1^1MC states as well as the S_6 (1^1MLCT) state are almost dark at both the CASSCF and CASPT2 levels of theory. The intensities of the bright S_{11} and S_{10} states are consistent at both the CASSCF and CASPT2 levels of theory. Accordingly, the oscillator strength of the 1^1IL (S_{11}) state is higher than the one of the 1^1MLCT (4^1B_2) state. This is in agreement with the experimental evidence, since higher intensities are obtained for the band at 3.96 than for the one peaking at ca. 4.13–4.34 eV.

The previous analysis serves to focus on some relevant states (the 4^1B_2 and 3^1A_1 MLCT states, the 5^1B_2 IL state, and exemplarily one MC state (1^1B_1)) for which the CASPT2/CASSCF results will be compared with different TD-DFT and RASPT2/RASSCF protocols in the coming subsections. However, before proceeding to discuss such calculations, we have considered of interest to analyze the SO effects on the spectroscopic properties of complex **1**. The SO-MS-CASPT2(14,13) values are also tabulated in Table 1 and illustrate that energetic shifts are present in all the low-lying excited states of complex **1**. The biggest shift is found in the case of the 2^1B_1 state (1^1MC state), amounting up to 0.04 eV as compared to the spin-free MS-CASPT2(14,13) value. The resulting SO state is mixed with a close-lying triplet excited state (the spin-free contributions of the resultant SO state are 72% of 2^1B_1 and 25% of 2^3A_2 ; the lowest lying triplet excited states are summarized in Table S1 of the Supporting Information). Such strong mixing is expected in states with participation of the ruthenium center, such as the

1^1MC states. The SO shifts obtained here are comparable to those obtained in other TM complexes, e.g., Os complexes.⁴⁴ In summary, the SO effects modify the spectroscopic properties of complex **1** but not significantly. SO effects are on the other hand indispensable to the interpretation of many other photo-physical phenomena, such as intersystem crossing (ISC) rates. We note that very large SO couplings (ca. 280 cm⁻¹) are found between the 3^1A_1 and T_1 (1^3A_2) states, which lie almost degenerate in energy favoring the ISC.

3.2. TD-DFT Calculations. The results obtained with different functionals for the excited states specified above are collected in Table 2. The values including CH₃CN are given in italics. The 1^1MC state (1^1B_1 state, following the CASSCF label of Table 1) is predicted within all the employed functionals in a range of ca. 0.4 eV, being rather immune to solvent effects. It has been previously seen that TD-DFT can succeed to describe d–d transitions of closed-shell TM carbonyl complexes,⁴⁵ such as Cr(CO)₆.⁴⁶ On the other hand, as pointed by Neese and co-workers, TD-DFT has drawbacks when describing d–d transitions in problematic configurations like d⁵ situations; for example, errors exceeding 0.6 eV were found in the TD-DFT energies of the [Ni(H₂O)₆]²⁺ complex (d⁸ configuration) with the B3LYP functional.¹⁶ TD-B3LYP was even not capable to reproduce the correct number of d–d excitations in this complex because some of them contained substantial double excitation character that TD-DFT cannot handle within the constraint of the adiabatic approximation. In our closed-shell complex **1**, meta-hybrids (namely the M06 and M06-2X) tend to slightly underestimate the d–d transition energy, especially M06, which yields the highest deviation (ca. 0.3–0.4 eV) in comparison to the rest of TD-DFT values and the MS-CASPT2(14,13) results. The range-separated functionals yield the larger values for the 1^1MC state, especially the LC- ω PBE functional which predicts 3.50 eV in good agreement with the CASPT2 value. Both hybrid and pure functionals predict the energy of the 1^1B_1 state in a small range. We note that the larger the percentage of exact HF-exchange is contained in the hybrid functionals, the higher are the excitation energies corresponding to the d–d transition; thus, B3LYP-35 (35% of exact exchange), PBE0 (25%), and B3LYP (20%) yield values of 3.34, 3.31, and 3.25 eV, in the gas phase, respectively. From all these values, we conclude that overall a reasonable description of the 1^1MC states

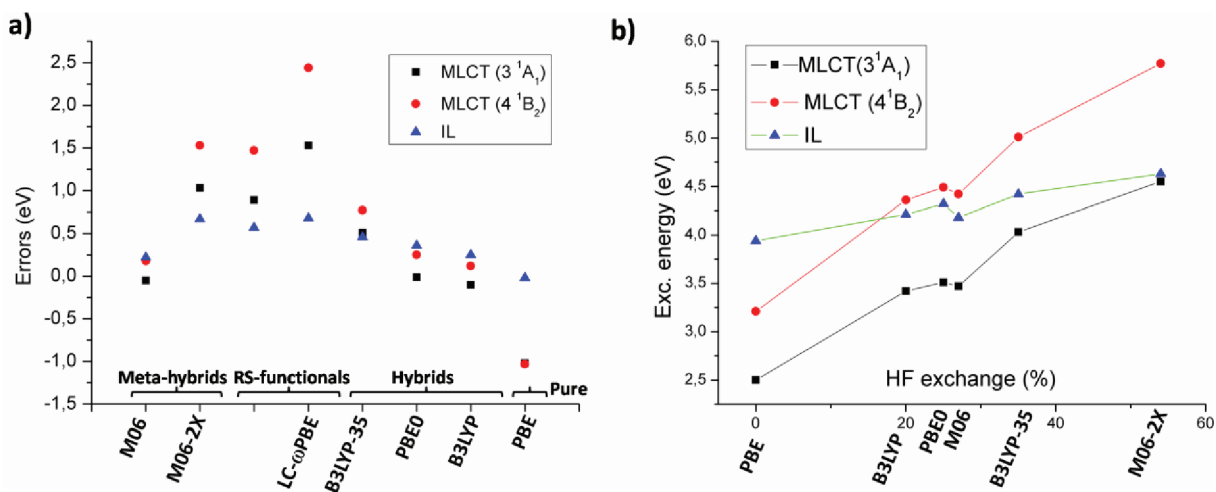


Figure 3. (a) PCM-TD-DFT errors (eV) in the electronic excitation energies of complex 1 for different types of excited states. (b) PCM-TD-DFT vertical excitation energies versus percentage of HF exchange of the density functionals for different types of excited states.

can be obtained for complex 1 with TD-DFT, and therefore we expect a similar success for 1MC states in other related closed-shell Ru(II) polypyridyl complexes.

Regarding the 1MLCT states of complex 1, namely the 3^1A_1 and 4^1B_2 states, one can clearly see at first sight the failure of some functionals behavior, which is well-known in CT states of organic dyes.⁴⁷ These states are specially subject of changing in the presence of a polar solvent. Indeed, both 1MLCT states are blue shifted by ca. 1 eV as compared to the gas phase values. It is worth noting that this effect manifests regardless the employed functional. Therefore, only the PCM-TD-DFT data will be compared with the experiment. The pure PBE functional leads to underestimations exceeding 1 eV for both 1MLCT states. An adequate description of 1MLCT states relies heavily on the choice of the functional. Conventional hybrid functionals have a clear tendency to increase the excitation energy of the 1MLCT states with increasing exact exchange. For instance, we see that the 3^1A_1 state is predicted at 3.42, 3.51, and 4.03 eV with B3LYP, PBE0, and B3LYP-35, respectively. These excitation energies are in good agreement with the experimental evidence, especially the B3LYP and PBE0 values. The best result for the 1MLCT states is obtained with the functional bearing an intermediate amount of exact exchange, i.e., PBE0, which delivers a deviation from the experiment of only 0.01 eV for the 3^1A_1 state. Indeed, 1MLCT states have been found to be best described with this functional for similar Ru complexes.⁴⁸ The trend of larger errors with increasing amount of exact exchange is also observed in the meta-hybrid functional M06 and M06-2X. The 3^1A_1 state is best described with the M06 functional (27% of exact exchange) which delivers a value of 3.47 eV (error 0.05 eV). The recently developed range-separated functionals employed here perform differently for describing the 1MLCT states of complex 1. LC- ω PBE hugely overestimates the excitation energies of the 1MLCT states (e.g., the 4^1B_2 state is overestimated by more than 2 eV). Interestingly, we note that CAM-B3LYP performs worse than the hybrid functionals B3LYP-35, PBE0, B3LYP, and M06. The 3^1A_1 state is theoretically predicted at 4.41 eV, and hence the error with respect to the experiment amounts almost to 1 eV. The 4^1B_2 state is also overestimated by 1 eV (see Table 2). This evidence is contrary to the good performance of CAM-B3LYP in the treatment of CT states of organic dyes.⁴⁹

But indeed, the agreement of CAM-B3LYP to describe MLCT states of other TM complexes^{21d,50} is not better. In any case, it is fair noting that the errors coming from different sources, as e.g., the selection of the exchange–correlation functional or the solvation method, might cancel among them leading to inconclusive results. We are confident, nevertheless, that the PCM method, despite its limitations, performs well in this complex.

The energies of the 1IL state are not as strongly dependent on the functional as the 1MLCT states, yielding values in the gas phase for the 3^1B_2 state that range from 4.82 eV with the LC- ω PBE functional to 4.02 eV with the pure PBE functional. We note that, oppositely to the 1MLCT states, inclusion of solvent effects within the limits of the PCM model leads only to red shifts of ca. 0.1 eV (in all the functionals, see Table 2). Among the hybrid functionals, the excitation energies are lower the smaller the amount of HF exchange, as it happened with the 1MC and 1MLCT states. Interestingly, and in concordance with the 1MLCT states, the agreement with the experiment is obtained with PBE, B3LYP, M06, or PBE0 (with errors for all the functionals well below 0.30 eV), rather than with functionals containing high percentage of exact exchange (as B3LYP-35 or M06-2X) or the range-separated functional CAM-B3LYP, even though the errors for the latter functional are in the acceptable range of accuracy of TD-DFT (ca. 0.57 eV). A similar behavior has been also observed for conjugated organic compounds, where the small maximum average errors in the description of local $\pi\pi^*$ excitations are obtained with the PBE0 functional.²⁰

Since the PCM-TD-DFT values look trustworthy, as long as a proper functional is chosen, we assign the UV–vis spectra of complex 1 in the following way: The weak band peaking at ca. 3.52 eV is due to the 3^1A_1 1MLCT state. The strong band peaking at 3.96 and 4.13–4.34 eV, with absorption intensities of 14 100 and 11 300–10 000 $M^{-1}cm^{-1}$, respectively, can be attributed to the 5^1B_2 (1IL) and the 4^1B_2 (1MLCT) states, respectively. Such assignment is done in view of the energetic order the last two states, which is reproduced with most of the TD-DFT flavors (in the presence of solvent) and the MS-CASPT2(14,13) calculations. The oscillator strengths of both states are also in accordance to this assignment. Thus, the 1IL state possesses higher oscillator strength than the 1MLCT state, in accordance to the experimental absorption intensities.

Table 3. Relative RASSCF($n,l,m;i,j,k$) and RASPT2($n,l,m;i,j,k$) Electronic Transitions Energies, ΔE (in eV), of the Main Electronic Excitations of Complex 1, at Different Levels of Theory, Compared to the SA-CASSCF and MS-CASPT2 and Available Experimental Data

States	SA-CAS (14,13)	MS-CASPT2 (14,13)	SA-RAS (16,4,4; 8,0,8) ^a	MS-RASPT2 (16,4,4; 8,0,8) ^b	SA-RAS ^a (16, l,m ; 2,11,3)			MS-RASPT2 ^b (16, l,m ; 2,11,3)			SA-RAS (16,2,2; 3,9,4) ^a	MS-RASPT2 (16,2,2; 3,9,4) ^b	SA-RAS (22,2,2; 6,9,6) ^a	MS-RASPT2 (22,2,2; 6,9,6) ^b	SA-RAS (16,3,3; 4,7,4) ^c	MS-RASPT2 (16,3,3; 4,7,4) ^d	Exptl
					l,m =2	l,m =3	$l=3$, $m=4$	l,m =2	l,m =3	$l=3$, $m=4$							
1^1B_1 (¹ MC)	3.59	3.49	4.50	3.36	3.99	4.03	4.04	3.31	3.31	3.31	4.20	3.23	4.53	3.55	5.10	3.36	-
3^1A_1 (¹ MLCT)	5.61	3.35	5.70	4.32	5.18	5.22	5.22	3.36	3.35	3.35	5.51	3.92	5.33	3.83	4.11	3.34	3.52 ^e 3.62 ^f
5^1B_2 (¹ IL)	6.64	4.49	6.41	4.83	6.08	6.07	6.07	4.83	4.85	4.87	6.83	4.19	6.62	4.30 ^g	7.07	3.70	3.96 ^e
4^1B_2 (¹ MLCT)	6.34	4.75	6.92	4.57	6.58	6.57	6.58	4.58	4.61	4.62	6.35	3.96	6.17	4.61 ^h	6.07	3.96	4.13 ^e
Number of CSF's	183150		73108		2230837 ($l,m=2$); 5495197 ($l,m=3$); 7399206 ($l=3, m=4$)						557036		3890064		390737		

^a SA-(3,3,4)-RASSCF($n,l,m;i,j,k$) calculations for A_1 , B_1 , and B_2 symmetries, respectively. ^b MS-(3,3,4)-RASPT2($n,l,m;i,j,k$) calculations for A_1 , B_1 , and B_2 symmetries, respectively. ^c SA-(3,3,5)-RASSCF($n,l,m;i,j,k$) calculations for A_1 , B_1 , and B_2 symmetries, respectively. ^d MS-(3,3,5)-RASPT2($n,l,m;i,j,k$) calculations for A_1 , B_1 , and B_2 symmetries, respectively. ^e In CH_3CN , from ref 23. ^f In CH_2Cl_2 , from ref 24. ^g These states are strongly mixed at the MS(4)-RASPT2(22,2,2;6,9,6) level of theory.

A comment on the performance of the different TD-DFT flavors to predict oscillator strengths is in order here. As expected and in agreement with the experiment, the intensity of the ¹MC state is very low, no matter which functional is used or whether solvation is included. Noteworthy, the oscillator strengths of the ¹MLCT states computed in the gas phase are underestimated with all the functionals in comparison to the PCM-TD-DFT values; very likely, this error is connected to the underestimation of the excitation energies, among other effects. The least robust excitation is the one corresponding to the intense ¹IL state, whose results are dependent on the functional employed. In gas phase, PBE and B3LYP predict oscillator strengths thrice smaller than the rest of functionals. The reason behind this fact might be the mixing of the intense ¹IL state with an $n\pi^*$ excitation, as reflected on the wave function coefficients. See, for example, that the $\pi\pi^*/n\pi^*$ excitation wave function coefficients in the ¹IL state with the B3LYP functional are 0.52/0.46, respectively, while with the PBE0 functional, such values are 0.65/0.20, respectively. A more uniform picture is obtained by comparing the PCM-TD-DFT oscillator strengths since all the functionals predict similar values for the ¹IL state.

In summary it can be seen that the different functionals examined here show a different performance on the calculations of the low-lying excited states of complex 1. Figure 3a displays the errors in the energy of the PCM-TD-DFT values of the ¹IL and the ¹MLCT states, taking as a reference the experimental values. A direct comparison between the oscillator strengths and the absorption intensities is not possible since some of the experimental bands overlap. Additionally, Figure 3b shows the excitation energies for relevant ¹MLCT and ¹IL states as a function of the percentage of HF exchange of the employed pure and hybrid/meta-hybrid functionals. There it can be clearly seen that higher energies are obtained with higher percentages of HF exchange, being the effect more pronounced for the MLCT states. Similar trends have been observed for other Ru complexes.⁴⁸ As discussed before, acute problems are found in the description of ¹MLCT states. These are only accurately calculated by hybrid functionals with intermediate percentages of exact exchange; the functionals M06, PBE0, and B3LYP, in this order, give the best accuracy when solvent effects

are considered. These three functionals plus the pure functional PBE seem to be the only ones which are able to get a reasonable value of the local ¹IL excitation of 1. Based on these results, the best balanced description of all kind of excited states of Ru(II)–polypyridyl complexes can be best obtained with M06 in first place and B3LYP and PBE0 in second and third places, respectively, even though slight underestimations of the d–d transition are found for the M06 functionals. The inclusion of solvent effects is mandatory, especially for the ¹MLCT states. Indeed, this combination (an hybrid functional with intermediate amount of exact exchange and the consideration of solvent effects via the PCM method) has been successfully used in other Ru(II) polypyridyl complexes.^{21b,c} In this sense, it is surprising that the CAM-B3LYP functional, designed a priori to deal with CT states does not improve the conventional hybrid functionals when describing both the ¹MLCT and ¹IL states. Very likely, the bad performance of CAM-B3LYP for describing ¹MLCT states is due to the short distance between the donor and acceptor moieties. In contrast to these conclusions, the pure PBE functional (also in combination with solvent effects) has been found to outperform the B3LYP functional in the description of the excited states of Fe(phen)₂(CN)₂.⁵¹ The reason behind the good performance of the pure functional PBE in this particular situation is probably the mixing of the d-based orbitals with the ligand-based orbitals, leading to an overlap of the orbitals involved in the state and hence to a loose CT character of these states. Another possible explanation might be cancellation effects, as we stated above. Whether this situation takes place in others TM complexes needs to be evaluated for each particular case.

3.3. RASPT2/RASSCF Calculations. The energies and oscillator strengths obtained with the different RASPT2/RASSCF protocols are compiled in Table 3. The computational strategies that we will discuss below have been chosen following two intentions: The first is to improve the quality of the previous CASPT2/CASSCF using an affordable configuration space. The second is to find general hints about how to face the treatment of excited states of related TM complexes at the RASPT2/RASSCF level of theory.

As a first step with respect the CAS(14,13) calculations, we include the $n_{Cl}-a_1$ orbital and two π^*_{CO} orbitals ($\pi^*_{CO}-b_1$ and

$\pi^*_{\text{CO}}-a_2$) in the active space. The simplest approach is to distribute all the orbitals within RAS1 and RAS3 subspaces and leave the RAS2 empty while allowing SDTQ excitations within the RAS1/RAS3 subspaces. This is denoted as RASSCF(16,4,4;8,0,8). As it can be seen in Table 3, the number of CSFs is heavily reduced in comparison to the CASSCF(14,13) calculations, by almost a factor of 3. Unfortunately, while cheap and easy, this procedure leads to poor results for both the low- and high-lying excited states. The 3^1A_1 state, calculated at 3.35 eV with CASPT2(14,13), is now predicted at 4.32 eV with RASPT2(16,4,4;8,0,8) and hence deviating 1 eV from the CASPT2 value. The RASPT2(16,4,4;8,0,8) relative energies of the high-lying 5^1B_2 and 4^1B_2 bright states are only about 0.3–0.2 eV different from the CASPT2(14,13) values but with shifts in different directions. An important difference between both methodologies is that, as a consequence of this reverse shift, the 1MLCT (4^1B_2) state is below the 1IL state (5^1B_2) at the RASPT2(16,4,4;8,0,8) level of theory. The latter result is suspicious in view of the previous CASPT2 and TD-DFT results. In principle, the solvent effects could reverse the order of the states, since the 1MLCT state is more sensible than the 1IL state to solvatochromic effects, as reflected with the PCM-TD-DFT calculations. Unfortunately, RASPT2 calculations in the presence of solvent effects are difficult and computationally too demanding to be investigated, so that no further conclusions can be reached. We note however that leaving the RAS2 empty has also given poor results in describing the singlet–triplet splitting of copper complexes.¹² Even though this strategy had resulted useful in describing ionization and low-lying excited states of simple systems, such as organic π conjugated compounds,¹⁰ the results obtained in this work also seem to confirm that in TM complexes better results are obtained if the RAS2 subspace is not empty.

Therefore, in the following we focus on finding the optimal composition of the RAS2 subspace, which, as we will show, is indeed the crucial step to obtain a balanced and accurate description of the relevant excited states of **1**. Moving the $n_{\text{Cl}}-a_1$, the $\pi_{\text{bpy}}-2a_2/\pi^*_{\text{bpy}}-2a_2$ pair and the two π^*_{CO} orbitals to the RAS1/RAS3 subspaces leads to the RASSCF(16, l,m ;2,11,3) calculations, which will allow up to D, T, or Q excitations, depending on the l and m indexes. As we see in Table 3, the introduction of these orbitals and therefore further correlation in the zeroth-order wave function implies a slightly better description of some states. Exemplarily, the 3^1A_1 states goes from 5.61 eV at CASSCF(14,13) to 5.18 eV at RASSCF(16,2,2;2,11,3) level. The perturbative treatment leads to very similar results: the 3^1A_1 state is predicted at 3.36 eV with RASPT2(16,2,2;2,11,3) level of theory, and the 5^1B_2 and 4^1B_2 states are now located at 4.83 and 4.58 eV, respectively. The results seem to be hardly improved with respect to the CASPT2(14,13) results, since the transition responsible for the lowest energy band is located at the same position and the error with respect to the peaks higher in energy (5^1B_2 and 4^1B_2) is now ca. 0.9 and 0.4 eV, respectively (obviating several effects, e.g., solvatochromic effects) at the RASPT2(16,2,2;2,11,3) level of theory. An improvement of these results can be attempted following two strategies: either increasing the allowed excitations within the RAS1/3 subspaces or changing the current RAS partition.

First, we shall discuss the former strategy. Allowing up to triple excitations is labeled by RASSCF(16,3,3;2,11,3), while up to triples–quadruples ($l = 3$ and $m = 4$) is denoted by RASSCF(16,3,4;2,11,3). As it can be seen, the number of CSFs for such approaches is over 5 and 7 million, respectively, on the

limit of computational feasibility. Disappointingly, as it can be seen in Table 3, the relative energies of all the states show negligible changes regardless the level of excitation, at both RASSCF and RASPT2 levels of theory. These calculations do highlight that an enormous number of CSFs does not guarantee the quality of the results. If the orbitals composing RAS2 subspace are not properly selected to achieve a reasonable reference space, then perturbation theory will not be able to get accurate results.

Therefore, we turn our attention to redefine the RAS partition. An additional pair of $\pi_{\text{bpy}}/\pi^*_{\text{bpy}}$ orbitals ($\pi_{\text{bpy}}-1b_1/\pi^*_{\text{bpy}}-1a_2$) with high and low occupation numbers, respectively, is moved from the RAS2 to the RAS1/3 subspaces. This strategy has the additional advantage that, in principle, it allows to include additional orbitals in the RAS1/3 subspaces because now the number of CSFs is considerably reduced. Allowing SD excitations is labeled as RASSCF(16,2,2;3,9,4) calculations. As it can be seen in Table 3, the number of CSF's is reduced by a factor of 10 in comparison to the RASSCF(16,2,2;2,11,3) calculations. Very interestingly, we note that the results obtained with the economic RASPT2(16,2,2;3,9,4) level of theory seem to be more balanced than the previous ones. The low-lying 3^1A_1 state is now predicted at 3.92 eV. If we consider the solvatochromic blue shift calculated with the PCM-TD-DFT model, this state deviates even more from the experiment than the previous RASPT2 schemes. On the other hand, we believe that the 5^1B_2 and 4^1B_2 states, located now at 4.19 and 3.96 eV, are in closer agreement with the experimental evidence. Thus, the 1IL state is only accurately described within this latter procedure (recall that solvatochromic effects are not so important for this state, as reflected by the PCM-TD-DFT calculations). Additionally, the 4^1B_2 1MLCT state is also better predicted at this level of theory in comparison to the previous RASPT2 results, assuming also a solvatochromic blue shift. The 1MC (1^1B_1) state seems to be rather immune to electronic correlation, and it is located at 3.23 eV, a very similar value as that obtained with the other RAS calculations. We note that the maximum deviations in the case of the 1MC state, among the different RASPT2 calculations, amount only to ca. 0.3 eV. This is due to the immunity of these states to static correlation effects. Indeed, deviations in the same order of eV have been reported analogously for different CASPT2 calculations in other TM complexes.³⁰ The RASPT2(16,2,2;3,9,4) oscillator strengths at this level of theory are reported in parenthesis in Table 1. In most of the cases, the oscillator strengths are in agreement with the MS-CASPT2(14,13) ones. The main discrepancy is found in the S_{10} (4^1B_2) state, where RASPT2(16,2,2;3,9,4) predicts more than 1 order of magnitude smaller oscillator strength. Since the gas phase TD-DFT results also point to a weak 4^1B_2 state, we are confident that the RASPT2(16,2,2;3,9,4) oscillator strengths are trustworthy.

Inclusion of further correlation orbitals leads to the RASPT2/RASSCF(22,2,2;6,9,6) calculations. With a few million of CSF's, these calculations are on the computational limit and would be prohibitive at the CASPT2/CASSCF level of theory. They are probably an illustration of the size of TM–polypyridyl complexes that can be nowadays calculated at the RASPT2/RASSCF level of theory. Moreover, and more interestingly, they allow evaluating whether all the correlation orbitals are necessary to get accurate results for complex **1**. As it can be seen in Table 3, the low-lying 3^1A_1 state is slightly better described than at the RASPT2(16,2,2;3,9,4) level of theory. The 1IL state is also

accurately predicted at the RASPT2(22,2,2;6,9,6) level of theory, and the ^1MC (1^1B_1) state seems to be (again) unaffected by the inclusion of further correlation effects. On the other hand, the $^1\text{MLCT}$ (4^1B_2) state is blue shifted as compared to the RASPT2(16,2,2;3,9,4) calculations, exhibiting similar values to the rest of CASPT2 and RASPT2 values. We note however that at this level of theory, the intense ^1IL (5^1B_2) and the $^1\text{MLCT}$ (4^1B_2) states are strongly mixed. For example, the wave function coefficients of the configuration state function obtained by the $4d_{yz}-a_2 \rightarrow \pi^*_{\text{bpy}}-1b_1/\pi_{\text{bpy}}-1a_2 \rightarrow \pi^*_{\text{bpy}}-2b_1$ excitations are 0.38/0.38, in the case of the latter state. As a consequence of this mixing (which we consider unlikely as compared to the rest of CASPT2 and RASPT2 calculations performed), the oscillator strengths computed at the alternative RASPT2(16,2,2;3,9,4) level of theory can be considered more accurate. Probably, a higher number of average states is necessary to describe correctly the B_2 states with this active space, but due to the expensive cost of these calculations, no further trials have been done. Therefore, when going from the RASPT2(16,2,2;3,9,4) to the RASPT2(22,2,2;6,9,6) calculations, almost no improvement in the relative energies is achieved. The more economic RASPT2(16,2,2;3,9,4) partition is more valuable here, giving intensities in closer agreement with the experiment and the PCM-TD-DFT values.

In view of the rather poor improvement when going from the RASPT2(16,2,2;3,9,4) to the RASPT2(22,2,2;6,9,6) calculations, we have again reformulated the partition of the RAS subspaces so that $\sigma-\pi$ correlation is primarily described with the RAS2. In the RASPT2/RASSCF(16,3,3;4,7,4) calculations, the RAS2 subspace is then composed by the correlation orbitals, and SDT excitations are allowed within the RAS1/3 orbitals. Noteworthy, such calculations are more economic than the RASSCF(16,2,2;3,9,4) ones, being the number of CSFs only twice that of the ones spanned by a (14,13) active space. As it can be seen in Table 3, the RASPT2 (16,3,3;4,7,4) results are the most accurate. The ^1IL and ^1MC are accurately described in comparison to the experiment and the TD-DFT calculations (recall that these states are rather immune to solvent effects at the PCM-TD-DFT level of theory). The $^1\text{MLCT}$ states are now predicted at 3.34 (3^1A_1) and 3.96 (4^1B_2) eV, yielding the closest results to the experiment among all the RASPT2 calculations even after the expected blue shift due to the solvent.

Summarizing, it seems that the main bottleneck to obtain a balanced description of the excited states of Ru(II) polypyridyl complexes is the RAS2 partition. Our RASPT2 (16,3,3;4,7,4) results indicate that to get spectroscopic accuracy, the correlation orbitals should be included into the RAS2, while those orbitals participating into the main excitations should be included into the RAS1/3. In contrast to these conclusions, Sauri and co-workers have reported that for the excited states of a free-base porphyrin, the highest accuracy is obtained when the MOs involved in the main excitations are included in the RAS2 subspace.¹¹ This advice can also be followed in TM complexes, as long as it is computationally feasible. In cases where this is not the case (which unfortunately can be many), our computations conclusively demonstrate that correlation orbitals involved in covalent metal–ligand bonds should be the ones first included into the RAS2. In the title molecule, this strategy, leading to the partition RASPT2 (16,3,3;4,7,4), also pays off in comparison with CASPT2(14,13), especially in the calculation of high-lying excited states, such as the 4^1B_2 state. There is an evident improvement in the description of the latter state achieved due

to the proper treatment of the very important static correlation effects related to covalent metal–ligand bonds; and most importantly, at almost no further computational expense. We truly believe that these strategies will open the study of other larger 4d, and even 5d, TM complexes with strong covalent metal–ligand bonds using the RASPT2 method. In complexes, such as $[\text{Ru}(\text{bpy})_3]^{2+}$, many MLCT, IL, and ligand-to-ligand CT transitions are found in the low-energy region as a consequence of the many low-lying π/π^* orbitals located on the ligands. Following our recipe, these ligand orbitals could be better allocated into the RAS1/3, leaving the RAS2 free for relevant correlation orbitals involved in covalent metal–ligand bonds. Only in this way these complexes, otherwise unreachable for CASPT2, might be faced with RASPT2.

4. CONCLUDING REMARKS

TM complexes are prototypes of systems where transitions of very different character, i.e., IL, MC, MLCT, and LMCT states, can be found. This makes them specially complicated to handle computationally so that balanced results can be obtained for all the excited states simultaneously. As an example of such a situation, we have calculated the excited states of the *trans*-(Cl)-Ru(bpy)Cl₂(CO)₂ complex with CASPT2/CASSCF and RASPT2/RASSCF protocols, allowing different partitions of the active space and different excitation levels. As found in the case of some organic and inorganic systems,¹¹ in order to get accurate results, the inclusion of correlation orbitals in the active space is important, however, much more significant seems to be the choice of the partition of the RAS subspaces, especially the RAS2. From the results obtained here, we conclude that at least for this type of Ru complexes, the RAS2 subspace should not be empty, but it must contain the correlation orbitals involved in the covalency of the metal–ligand bonds and only those. The orbitals involved in the main electronic excitations should be better allocated to the RAS1 and RAS3 subspaces. Once an optimal partition is achieved, SDT excitations within the RAS1/3 subspaces are sufficient to handle the additional dynamical correlation and thus obtain the right order of the states with accurate energies and intensities. These hints should be transferable to compute excited states of analogue complexes with strong covalent metal–ligand bonds, like the $[\text{Ru}(\text{bpy})_3]^{2+}$ complex, at the RASPT2/RASSCF level of theory. Needless to say, additional molecules should be studied before universal trends can be drawn.

The performance of several TD-DFT flavors is herein also assessed. Desirable is a general method which allows describing the different transitions contributing to the UV–vis spectrum. Solvent effects are found to be mandatory to obtain spectroscopic accuracy, especially in the case of MLCT states. We find that while MC transitions are rather robust to any of the functionals tested, MLCT states are only well described with functionals bearing intermediate amounts of exact exchange, such as M06, PBE0, and B3LYP, in combination with solvent effects. IL states are also best described with these functionals. In view of these conclusions, here we find that the best compromise to treat all the excited states of Ru(II)–polypyridyl complexes in a balanced manner is first M06, and then the B3LYP and PBE0 functionals. Further benchmark studies are required to establish general trends in TM spectroscopy.

This study clearly shows that the rationalization of the UV–vis spectra of TM complexes exclusively based on the matching of

experimental and theoretical TD-DFT bands might be dangerous without an initial exploration of the performance of different hybrid or range-separated functionals because some ¹MLCT states might be theoretically underestimated (by more than 1 eV in some cases) but match accidentally a different band of the experimental spectrum.

■ ASSOCIATED CONTENT

S Supporting Information. The relative SA-CASSCF-(14,13) and MS-CASPT2(14,13) lowest lying triplet excited states are shown in Table S1. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: leticia.gonzalez@univie.ac.at

Present Addresses

[†]Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1, 45470 Mülheim an der Ruhr, Germany.

[‡]Institute for Theoretical Chemistry, University of Vienna, Währingerstrasse 17, 1190 Vienna, Austria.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENT


This work has been funded by the Carl-Zeiss foundation (D.E.). Computer time in the Rechenzentrum of the Friedrich-Schiller-Universität Jena is gratefully acknowledged. We are grateful to S. Vancoillie for fruitful discussions and one of the referees for pointing to the solvatochromic shifts, which turned to be rather crucial to interpret the spectrum of *trans*(Cl)-Ru(bpy)Cl₂(CO)₂.

■ REFERENCES

- (1) (a) Gust, D.; Moore, T. A.; Moore, A. L. *Acc. Chem. Res.* **2009**, *42*, 1890–1898. (b) Rau, S.; Schäfer, B.; Gleich, D.; Anders, E.; Rudolph, M.; Friedrich, M.; Görls, H.; Henry, W.; Vos, J. G. *Angew. Chem., Int. Ed.* **2006**, *45*, 6215–6218. (c) Tschierlei, S.; Karnahl, M.; Presselt, M.; Dietzek, B.; Guthmüller, J.; González, L.; Schmitt, M.; Rau, S.; Popp, J. *Angew. Chem., Int. Ed.* **2010**, *122*, 4073–4076.
- (2) (a) Concepcion, J. J.; Jurss, J. W.; Brennaman, M. K.; Hoertz, P. G.; Patrocínio, A. O. T.; Iha, N. Y. M.; Templeton, J. L.; Meyer, T. J. *Acc. Chem. Res.* **2009**, *42*, 1945–1955. (b) Romain, S.; Vigara, L.; Llobet, A. *Acc. Chem. Res.* **2009**, *42*, 1944–1953.
- (3) (a) O'Regan, B.; Grätzel, M. *Nature* **1991**, *353*, 737. (b) Hagfeldt, A.; Grätzel, M. *Acc. Chem. Res.* **2000**, *33*, 269. (c) Grätzel, M. *Nature* **2001**, *414*, 338. (d) Grätzel, M. *Pure Appl. Chem.* **2001**, *73*, 459. (e) Benkö, G.; Kallioinen, J.; Korppi-Tommola, J. E. Y.; Yartsev, A. P.; Sundström, V. *J. Am. Chem. Soc.* **2002**, *124*, 489–493. (f) Kallioinen, J.; Benkö, G.; Myllyperkiö, P.; Khriachtchev, L.; Skärman, B.; Wallenberg, R.; Tuomikoski, M.; Korppi-Tommola, J.; Sundström, V.; Yartsev, A. P. *J. Phys. Chem. B* **2004**, *108*, 6365–6373. (g) Grätzel, M. *Inorg. Chem.* **2005**, *44*, 6841.
- (4) Andersson, K.; Malmqvist, P.-A.; Roos, B. O. *J. Chem. Phys.* **1992**, *96*, 1218.
- (5) González, L.; Escudero, D.; Serrano-Andrés, L. *Chem. Phys. Chem.* **2011**, DOI: 10.1002/cphc.201100200.
- (6) Pierloot, K.; Vancoillie, S. *J. Chem. Phys.* **2006**, *125*, 124303.
- (7) Gindensperger, E.; Köppel, H.; Daniel, C. *Chem. Comm.* **2010**, *46*, 8225–8227.
- (8) (a) Olsen, J.; Roos, B. O.; Jorgensen, P.; Jensen, H. J. A. *J. Chem. Phys.* **1988**, *89*, 2185. (b) Malmqvist, P.-A.; Rendell, A.; Roos, B. O. *J. Chem. Phys.* **1990**, *94*, 5477.
- (9) Malmqvist, P.-Å.; Pierloot, K.; Shahi, A. R. M.; Cramer, J. C.; Gagliardi, L. *J. Chem. Phys.* **2008**, *128*, 204109.
- (10) Shahi, A. R. M.; Cramer, C. J.; Gagliardi, L. *Phys. Chem. Chem. Phys.* **2009**, *11*, 10964–10972.
- (11) Sauri, V.; Serrano-Andrés, L.; Shahi, A. R. M.; Gagliardi, L.; Vancoillie, S.; Pierloot, K. *J. Chem. Theory Comput.* **2011**, *7*, 153–168.
- (12) Huber, S. M.; Shahi, A. R. M.; Aquilante, F.; Cramer, C. J.; Gagliardi, L. *J. Chem. Theory Comput.* **2009**, *5*, 2967–2976.
- (13) Casida, M. E. *Recent advances in Density Functional Methods. Part I*; World Scientific: Singapore, 1995.
- (14) Cramer, C. J.; Truhlar, D. G. *Phys. Chem. Chem. Phys.* **2009**, *11*, 10757–10816.
- (15) (a) Dreuw, A.; Weisman, J. L.; Head-Gordon, M. *J. Chem. Phys.* **2003**, *119*, 2943–2946. (b) Dreuw, A.; Head-Gordon, M. *Chem. Rev.* **2005**, *105*, 4009–4037.
- (16) Neese, F.; Petrenko, T.; Ganyushin, D.; Olbrich, G. *Coord. Chem. Rev.* **2007**, *251*, 288–327.
- (17) (a) Ernzerhof, M.; Scuseria, G. E. *J. Chem. Phys.* **1999**, *110*, 5029–5036. (b) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.
- (18) Vydrov, O. A.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 234109.
- (19) (a) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51–56. (b) Peach, M. J. G.; Benfield, P.; Helgaker, T.; Tozer, D. J. *J. Chem. Phys.* **2008**, *128*, 044118.
- (20) (a) Jacquemin, D.; Perpète, E. A.; Scuseria, G. E.; Ciofini, I.; Adamo, C. *J. Chem. Theory Comput.* **2008**, *4*, 123–135. (b) Escudero, D.; Trupp, S.; Bussemer, B.; Mohr, G.; González, L. *J. Chem. Theory Comput.* **2011**, *7*, 1062. (c) Jacquemin, D.; Preat, P.; Wathelet, V.; Fontaine, M.; Perpète, E. A. *J. Am. Chem. Soc.* **2006**, *128*, 2072.
- (21) (a) Bossert, J.; Daniel, C. *Coord. Chem. Rev.* **2008**, *23–24*, 2493–2503. (b) Happ, B.; Escudero, D.; Hager, M. D.; Friebe, C.; Winter, A.; Görls, H.; Altuntas, E.; González, L.; Schubert, U. S. *J. Org. Chem.* **2010**, *75*, 4025–4038. (c) Schulze, B.; Escudero, D.; Friebe, C.; Siebert, R.; Görls, H.; Köhn, U.; Altuntas, E.; Baumgärtel, A.; Hager, M. D.; Winter, A.; Dietzek, B.; Popp, J.; González, L.; Schubert, U. S. *Chem. Eur. J.* **2011**, *17*, 5494. (d) Guthmüller, J.; González, L. *Phys. Chem. Chem. Phys.* **2010**, *12*, 14812–14821. (e) Vlcek, A., Jr.; Zalis, S. *Coord. Chem. Rev.* **2007**, *251*, 258–287.
- (22) (a) Petit, L.; Maldivi, P.; Adamo, C. *J. Chem. Theory Comput.* **2005**, *1*, 953–962. (b) Holland, J. P.; Green, J. C. *J. Comput. Chem.* **2010**, *31*, 1008–1014.
- (23) Chardon-Noblat, S.; Deronzier, A.; Ziessel, R.; Zsoldos, D. *Inorg. Chem.* **1997**, *36*, 5384–5389.
- (24) Eskelinen, E.; Haukka, M.; Venäläinen, T.; Pakkanen, T. A.; Wasberg, M.; Chardon-Noblat, S.; Deronzier, A. *Organometallics* **2000**, *19*, 163–169.
- (25) Andrae, D.; Häusermann, U.; Dolg, M.; Stoll, H.; Preuss, H. *Theor. Chim. Acta* **1990**, *77*, 123–141.
- (26) (a) Pierloot, K.; Dumez, B.; Widmark, P.-O.; Roos, B. O. *Theor. Chim. Acta* **1995**, *90*, 87. (b) Roos, B. O.; Lindh, R.; Malmqvist, P.-Å.; Veryazov, A.; Widmark, P.-O. *J. Phys. Chem. A* **2005**, *109*, 6575–6579.
- (27) (a) Hess, B. A. *Phys. Rev. A* **1985**, *32*, 756. (b) Hess, B. A. *Phys. Rev. A* **1986**, *33*, 3742. (c) Jansen, G.; Hess, B. A. *Phys. Rev. A* **1989**, *39*, 6016.
- (28) (a) Aquilante, F.; Pedersen, T. B.; Lindh, R.; Roos, B. O.; De Meras, A. S.; Koch, H. *J. Chem. Phys.* **2008**, *129*, 8. (b) Aquilante, F.; Malmqvist, P. A.; Pedersen, T. B.; Ghosh, A.; Roos, B. O. *J. Chem. Theory Comput.* **2008**, *4*, 694.
- (29) Roos, B. O.; Andersson, K.; Fülscher, M. P.; Malmqvist, P.-Å.; Serrano-Andrés, L.; Pierloot, K.; Merchán, M. *Multiconfigurational Perturbation Theory: Applications in 35 Electronic Spectroscopy. In Advances in Chemical Physics: New Methods in Computational Quantum Mechanics*; Prigogine, I., Rice, S. A., Eds.; John Wiley & Sons: New York, 1996; Vol. XCIII, pp 219–332.
- (30) Pierloot, K. *Mol. Phys.* **2003**, *101*, 2083–2094.

- (31) Pierloot, K. Calculations of electronic spectra of transition metal complexes. In *Computational Photochemistry*; Olivucci, M., Eds.; Elsevier B. V.: Amsterdam, The Netherlands, 2005; pp 279–315.
- (32) Finley, J.; Malmqvist, P.-Å.; Roos, B. O.; Serrano-Andrés, L. *Chem. Phys. Lett.* **1998**, *288*, 299.
- (33) Malmqvist, P.-Å.; Roos, B. O. *Chem. Phys. Lett.* **1995**, *245*, 215.
- (34) Ghigo, G.; Roos, B. O.; Malmqvist, P.-Å. *Chem. Phys. Lett.* **2004**, *396*, 142.
- (35) Roos, B. O.; Andersson, K. *Chem. Phys. Lett.* **1995**, *245*, 215.
- (36) Roos, B. O.; Malmqvist, P.-Å. *Phys. Chem. Chem. Phys.* **2004**, *6*, 2919.
- (37) (a) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652. (b) Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.
- (38) Menucci, B.; Cappelli, C.; Guido, C. A.; Cammi, R.; Tomasi, J. *J. Chem. Phys. A* **2009**, *113*, 3009–3020.
- (39) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *215*, 241.
- (40) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- (41) (a) Cossi, M.; Barone, V.; Menucci, B.; Tomasi, J. *Chem. Phys. Lett.* **1998**, *286*, 253. (b) Menucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *106*, 5151.
- (42) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, revision A.1; Gaussian, Inc.: Wallingford, CT, 2009.
- (43) (a) Karlström, G.; Lindh, R.; Malmqvist, P.-Å.; Roos, B. O.; Ryde, U.; Veryazov, V.; Widmark, P.-O.; Cossi, M.; Schimmelpfennig, P.; Neogrady, P.; Seijo, L. *Comput. Mater. Sci.* **2003**, *28*, 222. (b) Aquilante, F.; De Vico, L.; Ferré, N.; Ghigo, G.; Malmqvist, P.-Å.; Neogrady, P.; Pedersen, T. B.; Pitonak, M.; Reiher, M.; Roos, B. O.; Serrano-Andrés, L.; Urban, M.; Veryazov, V.; Lindh, R. *J. Comput. Chem.* **2010**, *31*, 224.
- (44) Vallet, V.; Strich, A.; Daniel, D. *Chem. Phys.* **2005**, *311*, 13.
- (45) Daniel, C. *Coord. Chem. Rev.* **2003**, *238–239*, 143–146.
- (46) Rosa, A.; Baerends, E. J.; Gisbergen, S. J. A.; Lenthe, E. V.; Groeneveld, J. A.; Snijders, J. G. *J. Am. Chem. Soc.* **1999**, *121*, 10356.
- (47) (a) Serrano-Andrés, L.; Fülscher, M. P. *J. Am. Chem. Soc.* **1998**, *120*, 10912. (b) Peach, M. J. G.; Benfield, P.; Helgaker, T.; Tozer, D. J. *J. Chem. Phys.* **2008**, *128*, 044118.
- (48) (a) Záliš, S.; Ben Amor, N.; Daniel, C. *Inorg. Chem.* **2004**, *43*, 7978. (b) Ben Amor, N.; Záliš, S.; Daniel, C. *Int. J. Quantum Chem.* **2006**, *106*, 2458.
- (49) (a) Peach, M. J. G.; Ruth Le Sueur, C.; Ruud, K.; Guillaume, M.; Tozer, D. J. *Phys. Chem. Chem. Phys.* **2009**, *11*, 4465. (b) Plötner, J.; Tozer, D. J.; Dreuw, A. *J. Chem. Theory Comput.* **2010**, *6*, 2315–2324. (c) Jacquemin, D.; Perpète, A.; Scuseria, G. E.; Ciofini, I.; Adamo, A. *Chem. Phys. Lett.* **2008**, *465*, 226.
- (50) Fraser, M. G.; Blackman, A. G.; Irwin, G. I. S.; Easton, C. P.; Gordon, K. C. *Inorg. Chem.* **2010**, *49*, 5180.
- (51) Georgieva, I.; Aquino, A. J. A.; Trendafilova, N.; Santos, P. S.; Lischka, H. *Inorg. Chem.* **2010**, *49*, 1634–1646.

Predicting Nuclear Resonance Vibrational Spectra of [Fe(OEP)(NO)]

Qian Peng,[†] Jeffrey W. Pavlik,[†] W. Robert Scheidt,[†] and Olaf Wiest^{†,‡,*}[†]Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, Indiana 46556, United States[‡]School of Chemical Biology and Biotechnology, Peking University, Shenzhen Graduate School, Shenzhen 518055, China Supporting Information

ABSTRACT: Nuclear resonance vibrational spectroscopy (NRVS) is a sensitive vibrational probe for biologically important heme complexes. The exquisite sensitivity of the NRVS data to the electronic structure provides detailed insights into the nature of these interesting compounds but requires highly accurate computational methods for the mode assignments. To determine the best combinations of density functionals and basis sets, a series of benchmark DFT calculations on the previously characterized complex [Fe(OEP)NO] (OEP²⁻ = octaethylporphyrinato dianion) was performed. A test set of 21 methodology combinations including eight functionals (BP86, mPWPW91, B3LYP, PBE1PBE, M062X, M06L, LC-BP86, and ω B97X-D) and five basis set (VTZ, TZVP, and LanL2DZ for iron and 6-31G* and 6-31+G* for other atoms) was carried out to calculate electronic structures and vibrational frequencies. We also implemented the conversion of frequency calculations into orientation-selective mode composition factors (e^2), which can be used to simulate the vibrational density of states (VDOS) using Gaussian normal distribution functions. These use a series of user-friendly scripts for their application to NRVS. The structures as well as the isotropic and anisotropic NRVS of [Fe(OEP)NO] obtained with the M06L functional with a variety of basis sets are found to best reproduce the available experimental data, followed by B3LYP/LanL2DZ calculations. Other density functionals and basis sets do not produce the same level of accuracy. The noticeably worse agreement between theory and experiment for the out-of-plane NRVS compared with the excellent performance of the M06L functional for the in-plane prediction is attributed to deficiencies of the physical model rather than the computational methodology.

INTRODUCTION

Iron porphyrinates¹ are among the most important biological prosthetic groups and occur in many proteins and enzymes. A pivotal property of iron porphyrinates is the strong attraction of the central iron to axial ligands including histidine and diatomics like O₂ in hemoglobin (Hb) and myoglobin (Mb). The binding and dissociation reactions of small ligands like O₂ and NO in heme proteins are important biological processes.² In nature, NO is discriminated from O₂ quite efficiently, presumably due to conformational changes in the protein imparted upon ligand binding.³ Infrared and resonance Raman spectroscopy have provided insights into the interplay of structure and function of heme active sites.⁴ However, these techniques have some inherent limitations, especially in the low frequency regime where mode assignment is hampered by weak signals, spectral congestion, and low sensitivity to isotopic substitution.⁵ Nuclear resonance vibrational spectroscopy⁶ (NRVS) provides much higher selectivity because only the vibrational modes of the probe nucleus (⁵⁷Fe in the case of iron porphyrinates) contribute to the observed signal. Moreover, the NRVS intensity is directly related to the magnitude and direction of the motion; hence the method has a unique quantitative component in the measured vibrational spectrum. This method has been applied to heme enzymes, nitrogenase, and model complexes.⁷ However, the spectral crowding in the NRVS response region makes the spectra hard to identify for several vibrational modes, even for some very significant modes. In these cases, the computational prediction of NRVS spectra and comparison with experiment is an invaluable and indispensable tool in the assignment of the

observed modes. In turn, good agreement between experimental and computed NRVS spectra validates the computed results and increases the confidence in an analysis of the geometric and electronic structure of the entire molecule.

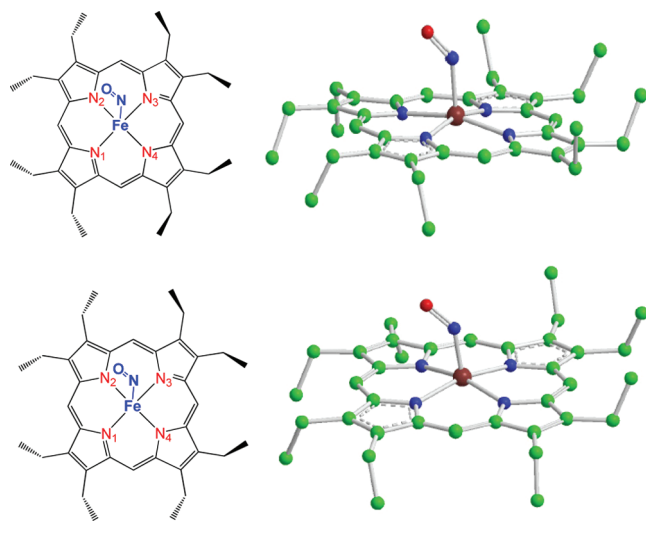
Density functional theory (DFT) methods now predict electronic structures and properties for molecules of increasing sizes, including detailed descriptions of their vibrational dynamics.⁸ Very recently, Noodleman and co-workers used a series of well-established density functionals to accurately calibrate the ⁵⁷Fe Mössbauer isomer shift and quadrupole splitting parameters.^{8d} Despite the complex electronic structure of heme complexes, DFT calculations are becoming increasingly accurate in the prediction of vibrational frequencies and are a very useful tool in mode assignments. The new generation of recently developed density functionals (such as the M0X series that address some of the shortcomings of previous DFT methods) holds significant promise for clarifying the character of a vibronic mode. The rich data set of vibrational frequencies, amplitudes, and directions available from NRVS can also provide a particularly rigorous test of the ability of DFT calculations to predict the vibrational dynamics of transition metal complexes.⁹

Lehnert and co-workers developed a useful method called “quantum chemistry centered normal coordinate analysis” (QCC-NCA)¹⁰ to fit some of the important NRVS peaks (bending and stretch modes) based on initial normal frequency calculations from Gaussian software. Clearly, a more rigorous and direct use of the

Received: September 15, 2011

Published: November 29, 2011

Scheme 1. Conformations of [Fe(OEP)(NO)]



normal modes would be preferable over such fitting procedures. In addition, the DFT methods applied to the prediction of NRVS have so far been limited to BP86 and B3LYP functionals, which do not describe dispersive interactions and suffer for incorrect descriptions of the self-interaction, especially in open shell systems.¹¹ As a result, they do not always have good agreement with experimental observations.^{5,12} To the best of our knowledge, there have been no systematic studies of more modern functionals that provide a much more balanced description of the electronic structure.

In this paper, we report a series of computational studies at different levels of theory, including the modern functionals that have not previously been tested for this purpose, with the goal of establishing best practices for the prediction and interpretation of NRVS data. Specifically, we compare the performance of different computational methods for (1) structure predictions, (2) the Fe–NO stretch and Fe–N–O bending vibrational modes, (3) the prediction of NRVS, and (4) directional NRVS (in-plane and out-of-plane). Finally, we discuss the effects of model issues. We also describe a set of user-friendly scripts that allow the direct conversion of frequency calculations into orientation-selective NRVS plots that can directly be compared to experimental data.

These calculations were benchmarked for [Fe(OEP)(NO)], a model complex for biologically relevant interactions of nitric oxide and heme iron for which there are high-resolution structures and NRVS data available. The rich data set of vibrational frequencies and directions available from NRVS for [Fe(OEP)(NO)] can provide a highly reliable test for our evaluation of DFT functionals and basis sets. At the same time, [Fe(OEP)(NO)] is a challenging molecule for DFT calculations because of the well-known difficulty in treating unpaired $S = 1/2$ spin systems.¹³ [Fe(OEP)NO] can adopt two conformations with different ethyl orientations in the solid state.¹⁴ One conformation is from a triclinic crystal that has four neighboring ethyl groups of OEP pointing to one face of the porphyrin, whereas the remaining four ethyl groups are in the opposite direction (Scheme 1, top). The other conformation is from a monoclinic crystal with five and three ethyl groups of OEP pointing to each face of the porphyrin plane (Scheme 1, bottom).

Ferrous heme-nitrosyls have low energy barriers for rotations of the NO ligand around the Fe–NO bond,¹⁵ which causes disorder in the NO orientation.¹⁶ The recent work by Lehnert

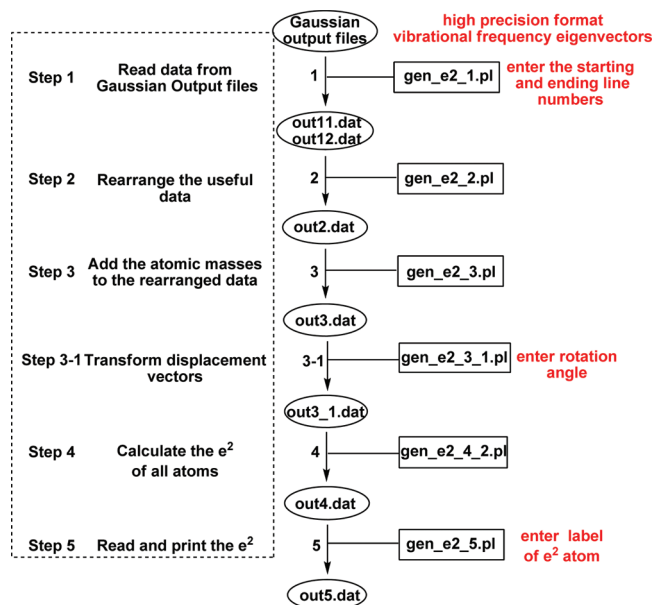


Figure 1. Flowchart of the scripts that generate e^2 in different orientations. The rectangles and ovals represent the script files and output files in every step, respectively.

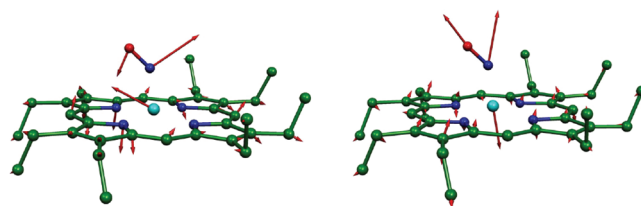


Figure 2. Fe–N–O bending (left) and Fe–NO stretch (right) vibrational modes. Hydrogen atoms and Fe–N bonds have been omitted for clarity. The vector is shown $100(m_j/m_{\text{Fe}})^{1/2}$ times longer than the zero-point vibrational amplitude of atom j .

and co-workers used B3LYP/LanL2DZ and BP86/LanL2DZ* calculations to test the 12 possible conformations for the disordered NO and rotated ethyl substituents of [Fe(OEP)(NO)].^{10d} Their prediction of NRVS spectra followed by QCC-NCA fitting has shown fair agreement with their powder measurements. However, comparisons of DFT predictions to powder measurements are lacking in the directional character of modes. We previously completed single-crystal measurements on [Fe(OEP)(NO)], taken at three orthogonal crystallographic directions, which shows significant directional anisotropy.^{12b} In contrast to typical nitrosyl iron porphyrins, [Fe(OEP)(NO)] exhibits a completely ordered NO group and is ideal for the oriented single-crystal NRVS experiment and for DFT calculations.

METHODS

Electronic Structure Calculations. The G09 program package¹⁷ was used to optimize the structures and for frequency analysis in our study. The model complex [Fe(OEP)(NO)] ($S = 1/2$) was fully optimized without any constraints using the spin unrestricted DFT method. The starting structure was obtained from the crystal structure of triclinic [Fe(OEP)(NO)].¹⁴ Frequency calculations were performed on the fully optimized structures at the

Table 1. Observed and Calculated Geometric Parameters of [Fe(OEP)(NO)]^a

[Fe(OEP)(NO)]		geometric parameters (pm or degree)									
functional	basis set	Fe–NO	Δ _{Fe–NO}	N–O	Δ _{N–O}	∠Fe–N–O	Δ _{∠Fe–N–O}	Fe–N _{short} ^c	Fe–N _{long} ^c	Δ _{long–short}	ref
triclinic crystal (experimental)		173.1		116.8		142.7		199.9	202.0	2.1	12b
BP86	6-31G*/VTZ	169.7	−3.4	119.2	2.4	143.2	0.5	200.3	203.1	2.8	12b
	6-31+G*(N,O) ^b /VTZ	169.6	−3.5	119.5	2.7	143.6	0.9	200.3	203.3	3.0	f
	6-31G*/TZVP	169.6	−3.5	119.1	2.3	144.9	2.2	200.9	203.9	3.0	f
MPWPW91	6-31G*/VTZ	169.8	−3.3	119.0	2.2	142.8	0.1	200.1	202.9	2.8	f
	6-31+G*(N,O) ^b /VTZ	169.6	−3.5	119.3	2.5	143.7	1.0	200.3	203.3	3.0	f
	6-31G*/TZVP	169.6	−3.5	118.9	2.1	144.7	2.0	200.8	203.8	3.0	f
B3LYP ^d	6-31G*/LanL2DZ	172.8	−0.3	117.0	0.2	139.8	−2.9	200.7	203.1	2.4	f
	LanL2DZ ^c	174.2	1.1	121.5	4.7	142.9	0.2	201.3	202.7	1.4	10d ^g
PBE1PBE	6-31G*/VTZ	170.7	−2.4	116.5	−0.3	140.2	−2.5	199.2	201.3	2.1	f
M062X ^d	LanL2DZ	231.1	58	118.2	1.4	125.8	−16.9	203.0	202.0	−1.0	f
M06L	6-31G*/VTZ	172.2	−0.9	118.0	1.2	140.2	−2.5	200.6	202.6	2.0	f
	6-31+G*(N,O) ^b /VTZ	172.4	−0.7	118.1	1.3	140.7	−2.0	200.8	203.1	2.3	f
	6-31G*/TZVP	172.0	−1.1	117.9	1.1	141.5	−1.2	201.1	203.2	2.1	f
	6-31+G*(N,O) ^b /TZVP	172.4	−0.7	118.1	1.3	140.9	−1.8	201.1	203.3	2.2	f
	6-31G*/LanL2DZ	172.2	−0.9	117.8	1.0	140.2	−2.5	200.2	202.7	2.5	f
	LanL2DZ ^c	173.5	0.4	121.9	5.1	141.9	−0.8	200.6	202.5	1.9	f
	LC-BP86	6-31G*/VTZ	170.9	−2.2	115.2	−1.6	139.1	−3.6	196.1	197.5	1.4
ωB97X-D	6-31G*/VTZ	172.6	−0.5	116.5	−0.3	139.3	−3.4	199.8	201.2	1.4	f
	6-31+G*(N,O) ^b /VTZ	172.3	−0.8	116.6	−0.2	139.7	−3.0	199.9	201.5	1.6	f

^a Basis sets: 6-31G* for H, C, N and O; VTZ, TZVP, or LanL2DZ for Fe. ^b Basis set: 6-31G* for H and C, 6-31+G* for N and O. ^c Basis set LanL2DZ for all of the atoms. ^d Basis set VTZ; TZVP for Fe is also applied but failed to obtain the converged wave functions. ^e Values for Fe–N_{short} and Fe–N_{long} (N_{short} = the average of two short Fe and porphyrin N atom, N_{long} = the average of two long Fe and porphyrin N atom; Δ_{long–short} is the difference between the experimental data and calculation with corresponding bond lengths and angles. ^f This work. ^g Only isotropic terms; for anisotropic terms, see this work.

same basis level to obtain the vibrational frequencies with the ⁵⁷Fe isotope set, which can yield inelastic scattering at the 14.4125 keV nuclear resonance line in the NRVS experiment.¹⁸ It is well-known that the frequencies obtained from harmonic frequency analyses are larger than the experimentally observed values due to the neglect of anharmonicity.¹⁹ This is typically addressed using scaling factors. However, the precise values are not only method and basis set dependent but are also different for different frequency regimes and have been validated mostly for pure organic molecules rather than the metal complexes discussed here. Therefore, the choice of the precise value would be ambiguous, and the frequencies reported here were not scaled. The frequency output data have been created using the high precision format vibrational frequency eigenvectors in order to calculate mode composition factors (e^2) and vibrational densities of state (VDOS) as described below.

We studied five classes of functionals: (1) generalized gradient approximation (GGA) functionals (BP86,²⁰ mPWPW91²¹) which contain the exchange and GGA correlation functionals; (2) hybrid-GGA functionals (B3LYP,²² PBE1PBE²³), which contain a mixture of Hartree–Fock exchange with DFT exchange-correlation; (3) hybrid meta-GGA functionals (M062X²⁴); (4) meta-GGA functionals (M06L²⁵), M06L being a local meta-GGA functional; and (5) long-range GGA functionals (LC-BP86,^{20,26} ωB97X-D²⁷) which contain long-range corrections. These were combined with different basis sets²⁸ as specified in the Results section. In general, we used triple-ζ valence basis sets with (TZVP) or without (VTZ) polarization functions or a double ζ effective core potential (LanL2DZ) on iron and 6-31G* or 6-31+G* basis sets for all other atoms. In order to allow comparison to Lehnert's results,^{10d} the LanL2DZ basis set was also tested for all atoms.

Calculation of the NRVS Data. The first step in the calculation of the NRVS data is the calculation of the predicted mode composition factors, which are based on the atomic displacements of each atom (r_i) from the analytical frequency analysis using DFT. The mode composition factors $e_{j\alpha}^2$, which represent the fraction of the kinetic energy in frequency mode α due to the motion of atom j ($j = ^{57}\text{Fe}$ for NRVS), provide a convenient quantitative comparison between measurements and calculations.²⁹ Mode composition factors are defined in eq 1:

$$e_{j\alpha}^2 = \frac{m_j r_j^2}{\sum m_i r_i^2} \quad (1)$$

where m_i is the atomic mass of atom i and r_i is the absolute length of the Cartesian displacement vector for atom i in Ångströms. The mode composition factors for different directions are defined in terms of an averaged porphyrin plane as in-plane, which can be calculated from a projection of the atomic displacement vector x and y (eq 2). The out-of-plane atomic displacement perpendicular to the resulting porphyrin plane for a normal mode is obtained from a projection of the atomic displacement vector z (eq 3).

$$e_{j\alpha, \text{in-plane}}^2 = \frac{m_j (r_{jx}^2 + r_{jy}^2)}{\sum m_i r_i^2} \quad (2)$$

$$e_{j\alpha, \text{out-plane}}^2 = \frac{m_j r_{jz}^2}{\sum m_i r_i^2} \quad (3)$$

The Perl scripts to calculate the mode composition factors are provided in the Supporting Information and directly read from a

Table 2. Observed and Calculated Vibrations of [Fe(OEP)(NO)]^a

[Fe(OEP)(NO)]		vibrational frequencies (cm ⁻¹)			
functional	basis set	$\nu(\text{Fe}-\text{NO})$	$\delta(\text{Fe}-\text{N}-\text{O})$	$\nu(\text{N}-\text{O})$	ref
triclinic crystal (experimental)		517	393	1673	12b
BP86	6-31G*/VTZ	623	417	1719	12b
	6-31+G*(N,O) ^b /VTZ	617	415	1689	e
	6-31G*/TZVP	618	410	1730	e
MPWPW91	6-31G*/VTZ	621	419	1727	e
	6-31+G*(N,O) ^b /VTZ	615	415	1698	e
	6-31G*/TZVP	616	410	1739	e
B3LYP ^d	6-31G*/Lanl2DZ	517/521	417	1828	e
	Lanl2DZ ^c	498/504	407	1616	10d ^f
PBE1PBE	6-31G*/VTZ	547	433	1885	e
M062X ^d	Lanl2DZ				e
M06L	6-31G*/VTZ	494/499	396	1798	e
	6-31+G*(N,O) ^b /VTZ	477/485	391	1779	e
	6-31G*/TZVP	495/500	388	1807	e
	6-31+G*(N,O) ^b /TZVP	477/485	386	1782	e
	6-31G*/Lanl2DZ	517	405	1804	e
	Lanl2DZ ^c	448	401	1624	e
LC-BP86	6-31G*/VTZ	551	443	2001	e
ω B97X-D	6-31G*/VTZ	532	425	1891	e
	6-31+G*(N,O) ^b /VTZ	522	422	1865	e

^a Basis sets: 6-31G* for H, C, N, and O; VTZ, TZVP, or Lanl2DZ for Fe. ^b Basis set: 6-31G* for H and C, 6-31+G* for N and O. ^c Basis set Lanl2DZ for all of the atoms. ^d Basis set VTZ; TZVP for Fe is also applied but failed to obtain the converged wave functions. ^e This work. ^f Only isotropic terms; for anisotropic terms, see this work.

G09 frequency output file using a high precision format for the vibrational frequency eigenvectors. Figure 1 showed the flow scheme for the scripts calculating $\epsilon_{j\alpha}^2$. The total procedures include six simple scripts, three of which (shown in red in Figure 1) need user input based on the specific molecule to be studied. In step 1, the starting and ending line numbers are needed for the frequency range of interest. In our case, a frequency range of 0–800 cm⁻¹ is defined to follow the range of NRVS in the experimental observations. In step 3-1, the orientation of interest is set as shown in Scheme 2. For porphyrins, the standard orientation used in G09 aligns the *x* and *y* coordinates along the iron–meso-carbon axes (4C-inplane), but the stretch vibration along the iron–nitrogen bond (4N-inplane) is needed. Thus, the calculation of $\epsilon_{j\alpha}^2$ requires in some cases a rotation of the coordinates by 45°. In the final step, the desired ϵ^2 data (in our case for ⁵⁷Fe, but available for any set of atoms) is read out.

The predicted mode composition factors $\epsilon_{j\alpha}^2$ can also be compared to the integrated spectral areas obtained from NRVS.⁵ Therefore, vibrational densities of state (VDOS) intensities can be simulated from the mode composition factors using the Gaussian normal distributions function, where the full width at half height (fwhh) is defined appropriately by considering the spectral resolution in the experiment. In this study, the MATLAB R2010a software was used to generate the predicted NRVS curves.

RESULTS AND DISCUSSION

Effects of DFT Method and Basis Set on the Calculated Geometry of [Fe(OEP)(NO)]. To investigate the effects of the DFT method and basis sets, we have performed DFT calculations on [Fe(OEP)(NO)] applying the functionals and basis sets discussed above. Tables 1 and 2 show the selected geometric and

vibrational properties of [Fe(OEP)(NO)] with the different basis sets in each density functional. In general, the basis set used is 6-31G* for H, C, N, and O and VTZ, TZVP, or Lanl2DZ for Fe. For selected functionals, a diffuse function 6-31+G* was added to N and O to better allow for molecular polarity and possible partial charge on the donor atoms. For the combination of the VTZ and TZVP basis sets on iron with the B3LYP and M062X functionals, strong spin contamination (0.87 for B3LYP and 0.95 for M062X) prevented, in some cases, the calculation of stable electronic structures.

As shown in Table 1, the BP86, mPWPW91, and PBE1PBE functionals, regardless of basis set, underestimate the Fe–NO bond length by up to 3.5 pm. The B3LYP, M06L, and ω B97X-D functionals reproduce the experimentally observed values quite accurately, while the M062X grossly overestimates the Fe–NO bond length. The basis set effect on this parameter is negligible. For the N–O bond length, the majority of the calculations with a variety of basis sets overestimate this parameter, but the accuracy is better than for the Fe–NO bond length. Interestingly, the basis set effect here is much higher, with the Lanl2DZ basis set on all atoms performing poorly compared to other methods. This can be rationalized by the iron back-donation to the N–O π^* orbital because this would make the Fe–NO bond short and elongate the N–O bond.

All functional/basis set combinations predict the Fe–N–O angle within 4° except for M062X, which gives a value 17° higher than the experimental value of 142.7°. The better of the predictions seem to occur in basis sets with simplified electronic systems such as the VTZ basis with no polarization function and Lanl2DZ with core effective potentials shown in Table 2 (BP86/VTZ, B3LYP/Lanl2dz, M06L/Lanl2dz).

An important feature of $[\text{Fe}(\text{OEP})(\text{NO})]$ that has been observed by the Scheidt group^{14b} is the different bond lengths of the four equatorial Fe–N bonds. Two short Fe–N_p distances are in the direction of the tilted NO ligand, while two long Fe–N_p distances are opposite the off-axis NO tilt. This anisotropic effect can be simplified to the different bond lengths of two short patterns and two long patterns, as shown in Table 1. All functional/basis set combinations except the M062X method predict this feature well. The observed difference of $N_{p\text{-long}} - N_{p\text{-short}}$ was 2.1 pm in the crystal structure and ranged from 1.0 to 3.0 pm in calculations.

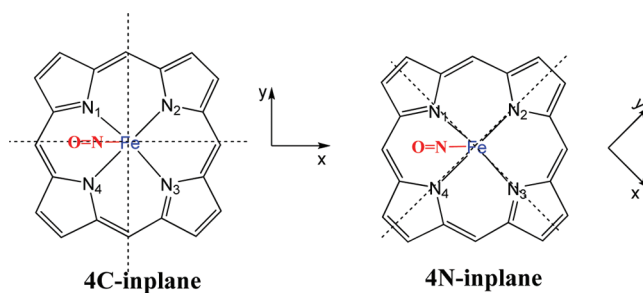
In summary, the M062X functional provides poor results for several of the geometric parameters, namely, the Fe–NO distance and the Fe–N–O angle, and cannot therefore be recommended for the systems under study here. Both of the GGA functionals, BP86 and mPWPW91, perform better but still underestimate the Fe–NO bond by about 3.5 pm and overestimate the N–O bond by about 2.5 pm. Better predictions were obtained from the B3LYP, M06L, and ω B97X-D functionals with a suitable basis set, all having agreement with experimental data within 1.0 pm for both bonds. M06L clearly stands out as providing the best agreement with experimental results.

Fe–NO Stretch and Fe–N–O Bending Vibrational Modes. The Fe–NO stretch and Fe–N–O bending modes are two major vibrations in $[\text{Fe}(\text{OEP})(\text{NO})]$ that have high frequencies and strong intensities. In the Fe–N–O bending mode, the motions of Fe and O are in the same direction and opposite to the nitrosyl nitrogen, while the Fe–NO stretch mode was predicted to have the opposite direction motion between the Fe and NO groups (Figure 2). Both the Fe–NO stretching and Fe–N–O bending frequencies are listed in Table 2 with the functional method and basis set designated.

As shown in Table 2, the GGA functionals, BP86 and mPWPW91, severely overestimate the Fe–NO stretching frequency ($>615 \text{ cm}^{-1}$ predicted vs 517 cm^{-1} observed). This difference is too large to be explained with anharmonicity and is likely because the Fe–NO bond lengths are underestimated. The other methods, including the hybrid-GGA, meta-GGA, and long-range GGA, provide better prediction for the stretch mode ranging from 448 cm^{-1} to 551 cm^{-1} . The exact frequency of 517 cm^{-1} was predicted by the B3LYP and M06L methods with the 6-31G*/lanl2dz basis set. The PBE1PBE, LC-BP86, and ω B97X-D were significantly less accurate than B3LYP and M06L. All of the Fe–N–O bending mode predictions are within 50 cm^{-1} of the experiment observation. The M06L method predicts it accurately to within $2\text{--}3 \text{ cm}^{-1}$ when the VTZ basis set for Fe was used. Although the BP86 and mPWPW91 provides the Fe–N–O angle closer to the experimental value, it does not give a better bending frequency prediction. It is clear that the harmonic approximation common to all calculations leads to deviations from the experimentally observed values but that the computed values still allow a clear assignment of the normal modes.

Prediction of Powder NRVS of $[\text{Fe}(\text{OEP})(\text{NO})]$. The main goal of this work was to determine the best practices for the prediction of NRVS. To investigate the effects of DFT methods and basis sets, we calculated the NRVS powder spectrum of $[\text{Fe}(\text{OEP})(\text{NO})]$. This spectrum has all of the vibrational modes of iron including the Fe–NO stretch mode and Fe–N–O bending modes discussed earlier. Selected NRVS predicted spectra are shown in Figure 3, and predicted spectra using other functionals can be found in the

Scheme 2. The Possible Orientations of Interest in Porphyrin Plane



Supporting Information. As can be seen in Figure 3, the NRVS can conveniently be divided into three frequency domains for which the mode assignment has been discussed in detail previously.^{12b,30} Here, we will discuss the performance of the different methods in terms of the different frequency domains.

In the region $>360 \text{ cm}^{-1}$, there are two important modes (Fe–NO stretch and Fe–N–O bend) in the observed NRVS spectrum and in each predicted spectrum (Figure 3A–F). There is an additional or partial peak in the $>360 \text{ cm}^{-1}$ region when the long-range LC-BP86 method was used (Figure 3C). It could be an overestimated mode from the $220\text{--}360 \text{ cm}^{-1}$ region. However, M06L was the best of the methods and was chosen to investigate the effects of the basis sets at the NRVS level (Figure 3D–F). As seen in Figure 3D and E, increasing the basis function from 6 to 31G* to 6-31+G* can move the Fe–NO stretch, and it is somewhat underestimated. In Figure 3D, the bending mode and stretch mode had shifted from 398 and 496 to 392 and 481, respectively. Similar trends can be seen in Figure 3E. The VTZ and TZVP basis sets show little difference. However, the Lanl2DZ basis set gives inconsistent frequency predictions (Figure 3F).

In the region $220\text{--}360 \text{ cm}^{-1}$, the two high-intensity peaks are observed with considerable spectral crowding. Most mode frequencies are poorly predicted by all functionals except M06L. The worst predictions are obtained by using GGA functionals with three apparently independent peaks (Figure 3A). The largely overestimated spectra are predicted with the long-range GGA functionals. As shown in Figure 3C, the major peaks move to the high-frequency region and are far away from the experimental observation designated by the black line. The better spectra are generated by the M06L functional with suitable basis set combinations. The spectra are almost coincident with the experimental spectra when the VTZ or TZVP basis set for Fe is used in the calculation (Figure 3D and E). However, the M06L functional is not quite as good when the Lanl2DZ basis set is used (Figure 3F).

All of the calculations gave similar spectral predictions in the $<220 \text{ cm}^{-1}$ region where the important doming mode occurs. However, all DFT methods used here underestimate the doming mode frequency and overestimate its intensity. The results of the calculations should therefore be interpreted with caution.

The In-Plane and Out-Plane NRVS of Single-Crystal $[\text{Fe}(\text{OEP})(\text{NO})]$. Comparisons of DFT predictions to powder measurements do not account for the directional character of modes—that is substantial. A more detailed comparison needs to be done to capture the directional anisotropy. In porphyrin chemistry, analysis confined to the averaged porphyrin plane is usually defined as in-plane, and two directions are defined (in-plane x and in-plane y). The third direction is

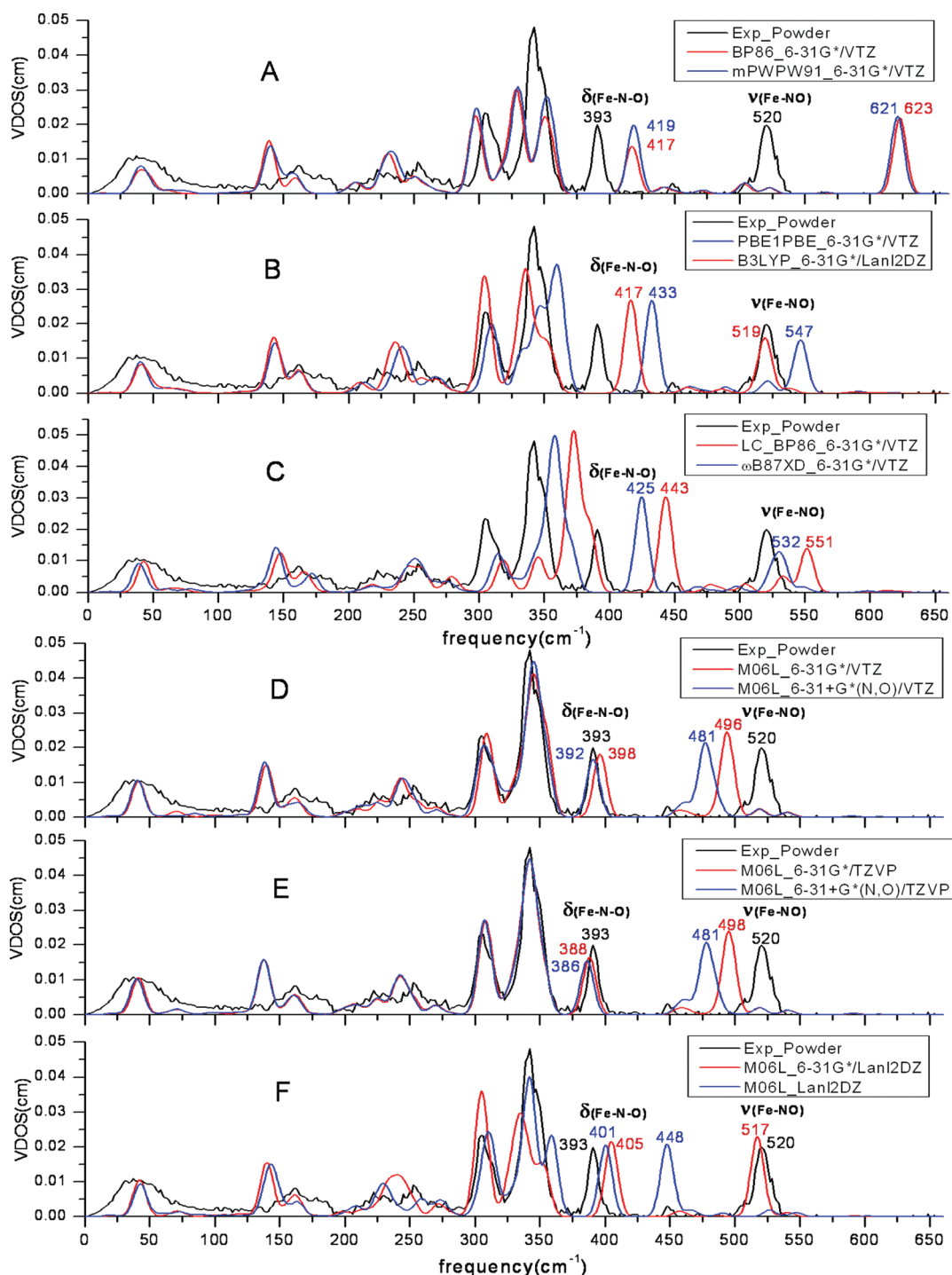


Figure 3. NRVS spectra of [Fe(OEP)(NO)] with a variety of methods and basis sets with 12 cm^{-1} fwhh. The experimental observation of [Fe(OEP)(NO)] powder is colored black, and the prediction data are colored red or blue (the same color strategy and fwhh in Figures 4–6).

perpendicular to the porphyrin plane and defined as out-of-plane or the z direction. In this paper, in-plane x is parallel to the Fe–NO plane and y is perpendicular to the Fe–NO plane, as shown in Scheme 2. This axis selection is called 4C-inplane. Only functionals that predicted the powder spectrum well are used in the comparisons of x – y – z -directional spectra (Figure 4–6). The published data with the BP86 method were chosen to highlight the superiority of the M06L functional. The orientation-selective spectra using

other functionals can also be found in the Supporting Information.

The possible anisotropy in the in-plane NRVS spectrum can be shown by a measurement in two orthogonal in-plane directions. Conveniently, we chose, for measurements and predictions, x to be along the projection of the FeNO plane and y to be orthogonal. These are shown in Figures 4 and 5, respectively. As shown in Figure 4, there are three observed peaks in the $>300 \text{ cm}^{-1}$ region, including one doublet peak at $300\text{--}325 \text{ cm}^{-1}$.

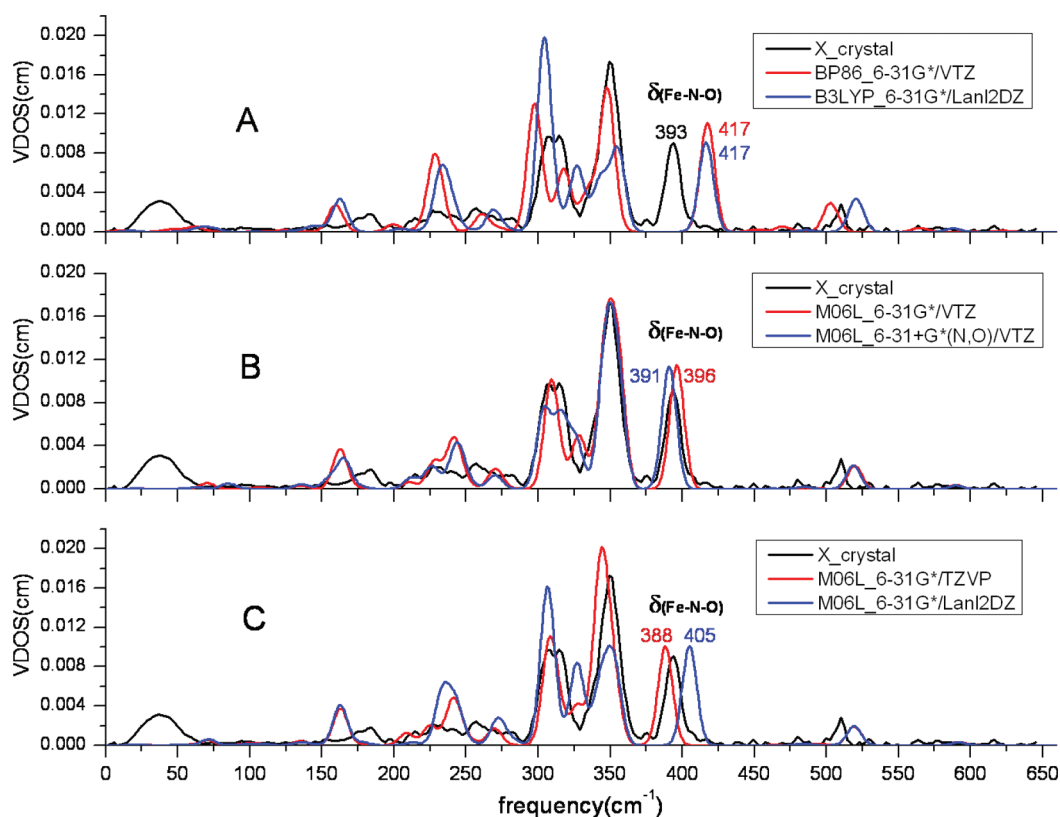


Figure 4. NRVs spectra of $[\text{Fe}(\text{OEP})(\text{NO})]$ with selected methods and basis sets in the x direction, which is parallel to the intersection of porphyrin and Fe–N–O plane.

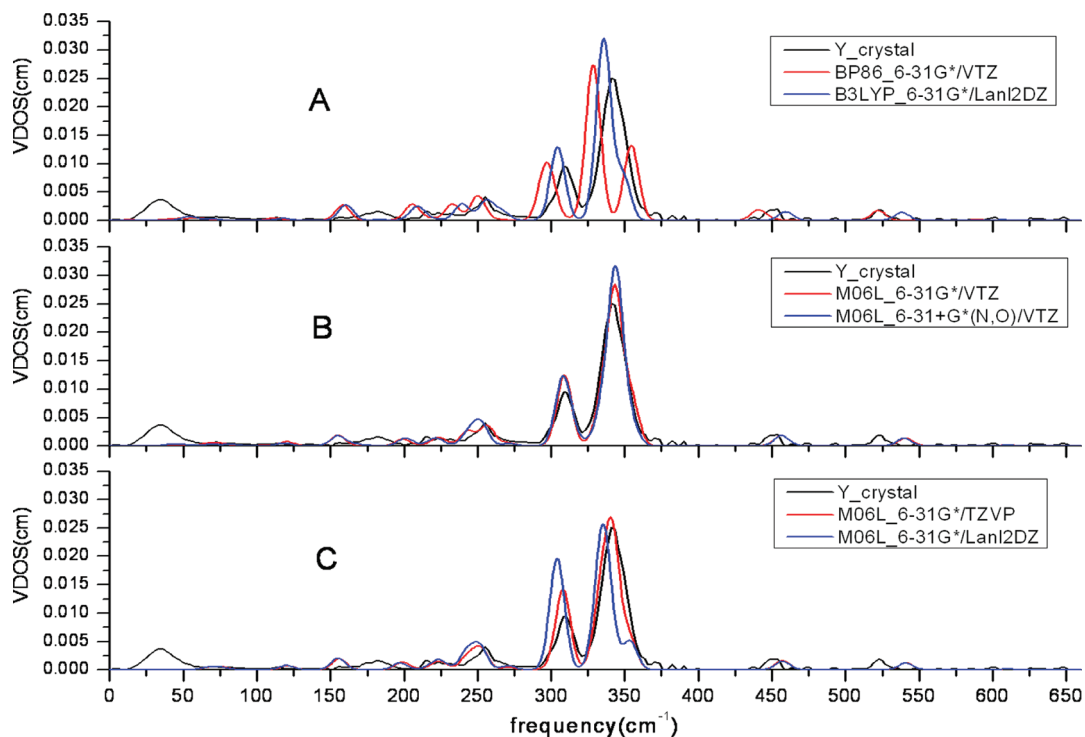


Figure 5. NRVs spectra of $[\text{Fe}(\text{OEP})(\text{NO})]$ with selected methods and basis sets in the y direction, which is perpendicular to the Fe–N–O plane.

The observed doublet peak is only accurately predicted with the diffused function 6-31+G* for N and O (Figure 4B). The Lan12DZ

basis set is unfavorable in the prediction of the NRVs spectrum, because more peaks are predicted from 300 cm^{-1} to 360 cm^{-1}

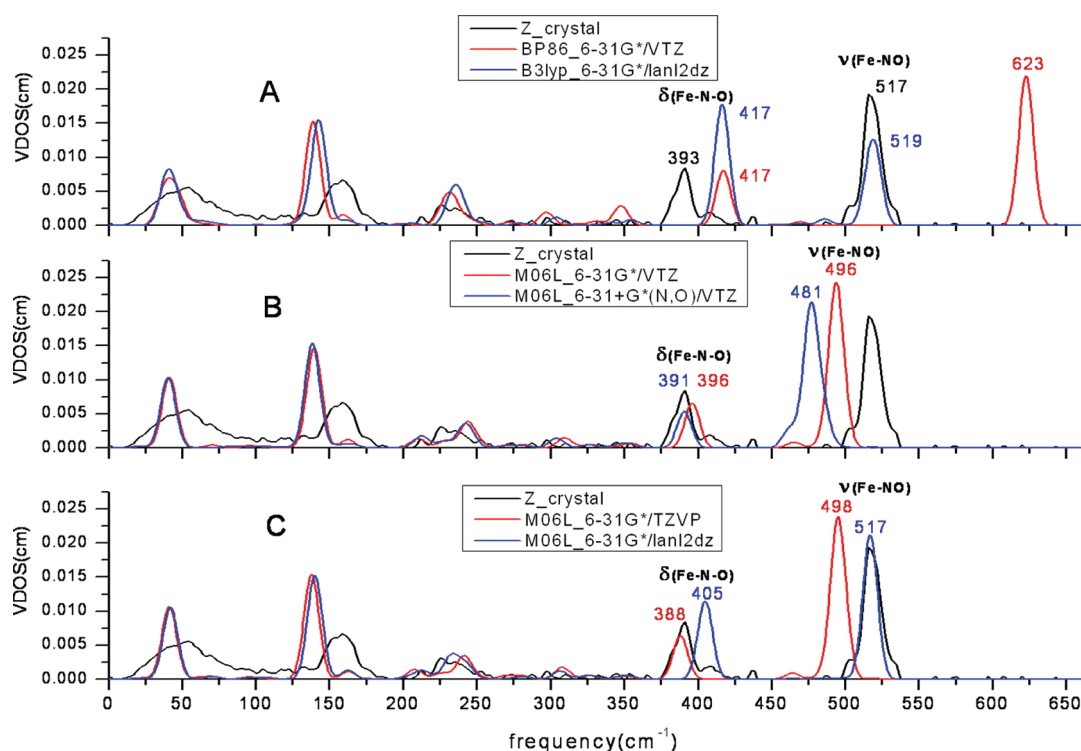


Figure 6. NRVS spectra of $[\text{Fe}(\text{OEP})(\text{NO})]$ with selected methods and basis sets in the z direction, which is perpendicular to the average porphyrin plane.

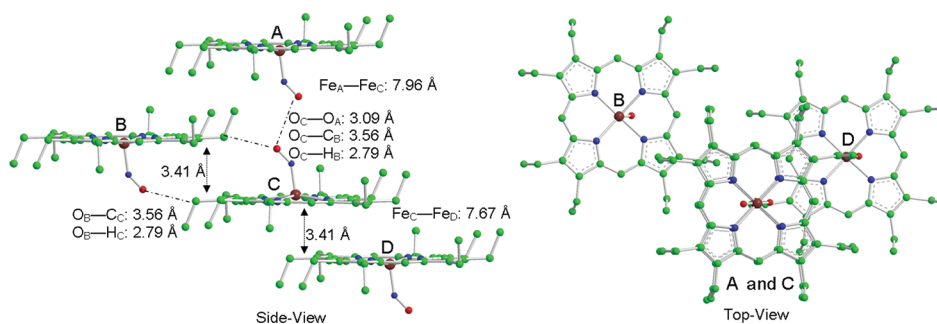


Figure 7. Crystal structures of triclinic $[\text{Fe}(\text{OEP})(\text{NO})]$. Hydrogen atoms have been omitted for clarity. Atom color: Fe (dark red), N (blue), O (red), C (green).

than observed and the bending mode around 390 cm^{-1} in the x direction is overestimated in both cases (Figure 4A/blue line and C/blue line). In sharp contrast, all of the major peaks in the NRVS along the y direction were underestimated when Lanl2DZ was used (Figure 5A/blue line and C/blue line). The BP86 functional is ruled out because three peaks are predicted, and only two are observed (Figure 5A). M06L with the VTZ or TZVP basis set shows better prediction than all others not only along the x direction but also along the y direction of NRVS.

The predictions of the out-of-plane NRVS spectra, shown in Figure 6, do not agree as well with experimental data as the in-plane spectra. The predicted doming modes are around 140 cm^{-1} regardless of the functional and basis set used, while the experimental observation is a broad peak at 160 cm^{-1} . The bending mode in the z -direction component is better predicted only when the M06L and VTZ basis sets were used, as shown in

Figure 6B. It is surprising that the stretch mode was perfectly predicted by the DFT calculation with the Lanl2DZ basis set, which in general gave poor predictions, as discussed above (Figure 6A and C). This presents a dilemma when using the Lanl2DZ basis set, because it predicts the stretch mode perfectly but fails on others modes. The diffuse function is not suitable for the prediction of the stretch mode, and it underestimates the experimental observations (Figure 6B). It is also inconsistent with the basis set strategy for in-plane NRVS.

A probable explanation is the difference between the gas-phase calculation and solid-state experiment. The predicted model is a single molecule without any intermolecular interactions, while both the experimental powder and crystal are in a crystal lattice and have intermolecular interactions or cooperativity. Figure 7 presents some likely intermolecular interactions. The distance between crystal layers is 3.41 Å , and the closest distance between two irons is 7.67 Å . This suggests that an intrinsic interaction

between different crystal layers could affect the out-of-plane modes more than in-plane ones. For the in-plane modes, the intermolecular close contacts cannot easily influence the iron center due to long distances. Therefore, the in-plane NRVS spectra may be more easily modeled. These unaccounted for forces can also explain why the out-of-plane spectra are always more difficult to predict accurately. The perfectly predicted stretching frequency by LanL2DZ is probably the result of chance rather than ideal modeling and basis set because it fails to predict other modes.

CONCLUSIONS

It is important to validate density functional methods in order to accurately predict directional anisotropy, a new feature in nuclear resonance vibrational spectroscopy (NRVS),^{12b} and to reliably assign the vibrational modes. The scripts discussed here and made available in the Supporting Information allow the facile calculation of orientation-selective mode composition factors (e^2), vibrational densities of states (VDOS), and NRVS data from standard Gaussian outputs. The extensive benchmarking of 21 different electronic structure methods for the representative case of the [Fe(OEP)NO] complex indicates that the M06L functional with suitable basis sets such as VTZ/6-31+G* or TZVP/6-31+G* provides the best agreement between calculated and experimental structures and NRVS data, followed by the B3LYP/6-31G*/LanL2DZ method. This is presumably due to the accurate description of the open-shell system and accurate excitation energies,³¹ which are likely to be important due to the low lying excited states in heme complexes, by the M06L functional.

A comparison of computational NRVS predictions and experimental data reveals that the M06L functional shows excellent agreement in the frequency domains of $>360\text{ cm}^{-1}$ and $200\text{--}360\text{ cm}^{-1}$. However, the frequencies of the modes in the region below 200 cm^{-1} are underestimated by all methods, including the M06L functional. A more detailed analysis of the anisotropic NRVS data shows that the M06L functional gives very good results for in-plane (x and y) but less so for out-of-plane (z) NRVS. The prediction of the in-plane NRVS using the M06L functional still exhibits the best performance, while the out-of-plane vibrations are less well predicted. This is most likely due to limitations of the model, which does not consider the crystal packing contacts that more influence the out-of-plane but not the in-plane vibrations.

In summary, the protocol presented here allows the facile and accurate prediction of NRVS data for the purpose of making assignments and to understand the detailed geometric and electronic structure of heme complexes.

ASSOCIATED CONTENT

S Supporting Information. Perl scripts, procedure for using the scripts, and e^2 values and computed NRVS spectra in three orthogonal directions as well as Cartesian coordinates and total energies of all structures discussed. This material is available free of charge via the Internet at <http://pubs.acs.org>

AUTHOR INFORMATION

Corresponding Author

*Tel.: +5746315876. Fax: +5746316652. E-mail: owiest@nd.edu

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

We gratefully acknowledge support of this research by the National Institutes of Health (Grant GM-38401 to W.R.S.). Generous allocation of computing resources by the Center for Research Computing at the University of Notre Dame and by the National Science Foundation through TeraGrid resources under grant number TG-CHE090124 are also acknowledged.

REFERENCES

- (1) (a) Berezin, B. D. *Coordination Compounds of Porphyrins and Phthalocyanines*; John Wiley & Sons Ltd.: New York, 1981; pp 1–10. (b) *Iron Porphyrins*; Lever, A. B. P., Gray, H. B., Eds.; Addison-Wesley Publishing Company Inc.: Reading, MA, 1983; pp 1–274.
- (2) Richter-Addo, G. B.; Legzdins, P.; Burstyn, J. *Chem. Rev.* **2002**, *102*, 857.
- (3) (a) Yu, A. E.; Hu, S.; Spiro, T. G.; Burstyn, J. N. *J. Am. Chem. Soc.* **1994**, *116*, 4117. (b) Negrerie, M.; Kruglik, S. G.; Lambry, J.; Vos, M. H.; Martin, J.; Franzen, S. *J. Biol. Chem.* **2006**, *281*, 10389.
- (4) Vogel, K. M.; Kozlowski, P. M.; Zgierski, M. Z.; Spiro, T. G. *Inorg. Chim. Acta* **2000**, *297*, 11.
- (5) Leu, B. M.; Zgierski, M. Z.; Wyllie, G. R. A.; Scheidt, W. R.; Sturhahn, W.; Alp, E. E.; Durbin, S. M.; Sage, J. T. *J. Am. Chem. Soc.* **2004**, *126*, 4211.
- (6) Scheidt, W. R.; Dubin, S. M.; Sage, J. T. *J. Inorg. Biochem.* **2005**, *99*, 60.
- (7) (a) Xiao, Y.; Wang, H.; George, S. J.; Smith, M. C.; Adams, M. W.; Jenney, F. E.; Sturhahn, W., Jr.; Alp, E. E.; Zhao, J.; Yoda, Y.; Dey, A.; Solomon, E. I.; Cramer, S. P. *J. Am. Chem. Soc.* **2005**, *127*, 14596. (b) Leu, B. M.; Zgierski, M. Z.; Wyllie, G. R.; Scheidt, W. R.; Sturhahn, W.; Alp, E. E.; Durbin, S. M.; Sage, J. T. *J. Am. Chem. Soc.* **2004**, *126*, 4211. (c) Achterhold, K.; Sturhahn, W.; Alp, E. E.; Parak, F. G. *Hyperfine Interact.* **2002**, *3*, 141–142. (d) Xiao, Y.; Tan, M. L.; Ichiye, T.; Wang, H.; Guo, Y.; Smith, M. C.; Meyer, J.; Sturhahn, W.; Alp, E. E.; Zhao, J.; Yoda, Y.; Cramer, S. P. *Biochemistry* **2008**, *47*, 6612. (e) Xiao, Y.; Fisher, K.; Smith, M. C.; Newton, W. E.; Case, D. A.; George, S. J.; Wang, H.; Sturhahn, W.; Alp, E. E.; Zhao, J.; Yoda, Y.; Cramer, S. P. *J. Am. Chem. Soc.* **2006**, *128*, 7608. (f) Rai, B. K.; Durbin, S. M.; Prohofsky, E. W.; Sage, J. T.; Wyllie, G. R.; Scheidt, W. R.; Sturhahn, W.; Alp, E. E. *Biophys. J.* **2002**, *82*, 2951. (g) Rai, B. K.; Durbin, S. M.; Prohofsky, E. W.; Sage, J. T.; Ellison, M. K.; Roth, A.; Scheidt, W. R.; Sturhahn, W.; Alp, E. E. *J. Am. Chem. Soc.* **2003**, *125*, 6927. (h) Rai, B. K.; Prohofsky, E. W.; Durbin, S. M. *J. Phys. Chem. B* **2005**, *109*, 18983. (i) Cramer, S.; Xiao, Y.; Wang, H.; Guo, Y.; Smith, M. *Hyperfine Interact.* **2007**, *170*, 47. (j) Petrenko, T.; DeBeer, S.; Aliaga-Alcalde, George, N.; Bill, E.; Mienert, B.; Xiao, Y.; Guo, Y.; Sturhahn, W.; Cramer, S. P.; Wieghardt, K.; Neese, F. *J. Am. Chem. Soc.* **2007**, *129*, 11053.
- (8) (a) Paulsen, H.; Winkler, H.; Trautwein, A. X.; Grünsteudel, H.; Rusanov, V.; Toftlund, H. *Phys. Rev. B* **1999**, *59*, 975. (b) Paulsen, H.; Benda, R.; Herta, C.; Schünemann, V.; Chumakov, A. I.; Duelund, L.; Winkler, H.; Toftlund, H.; Trautwein, A. X. *Phys. Rev. Lett.* **2001**, *86*, 1351. (c) Paulsen, H.; Rusanov, V.; Benda, R.; Herta, C.; Schünemann, V.; Janiak, C.; Dorn, T.; Chumakov, A. I.; Winkler, H.; Trautwein, A. X. *J. Am. Chem. Soc.* **2002**, *124*, 3007. (d) Sandala, G. M.; Hopmann, K. H.; Ghosh, A.; Noodleman, L. *J. Chem. Theory Comput.* **2011**, *7*, 3232.
- (9) (a) Zhou, M.; Andrews, L.; Bauschlicher, C. W., Jr. *Chem. Rev.* **2001**, *101*, 1931. (b) Ghosh, A.; Bocian, D. F. *J. Phys. Chem.* **1996**, *100*, 6363. (c) Kozlowski, P. M.; Jarzecki, A. A.; Pulay, P.; Li, X.-Y.; Zgierski, M. Z. *J. Phys. Chem.* **1996**, *100*, 13985. (d) Kozlowski, P. M.; Spiro, T. G.; Bérces, A.; Zgierski, M. Z. *J. Phys. Chem. B* **1998**, *102*, 2603. (e) Ghosh, A.; Skancke, A. *J. Phys. Chem. B* **1998**, *102*, 10087. (f) Kozlowski, P. M.; Spiro, T. G.; Zgierski, M. Z. *J. Phys. Chem. B* **2000**, *104*, 10659. (g) Cao, Z.; Hall, M. B. *J. Am. Chem. Soc.* **2001**, *123*, 3734. (h) Franzen, S. *J. Am. Chem. Soc.* **2001**, *123*, 12578. (i) Steene, E.; Wondimagegn, T.; Ghosh, A. *J. Inorg. Biochem.* **2002**, *88*, 113. (j) Ohta, T.; Matsuura, K.; Yoshizawa, K.; Morishima, I. *J. Inorg. Biochem.* **2000**, *82*, 141. (k) Maréchal, J.-D.; Maseras, F.; Lledós, A.; Mouawad, L.;

Perahia, D. *Chem. Phys. Lett.* **2002**, 353, 379. (l) Liao, M.-S.; Scheiner, S. *J. Chem. Phys.* **2002**, 116, 3635.

(10) For selected papers: (a) Praneeth, V. K. K.; Neather, C.; Peters, G.; Lehnert, N. *Inorg. Chem.* **2006**, 45, 2795. (b) Paulat, F.; Berto, T. C.; DeBeer George, S.; Goodrich, L.; Praneeth, V. K. K.; Sulok, C. D.; Lehnert, N. *Inorg. Chem.* **2008**, 47, 11449. (c) Berto, T. C.; Praneeth, V. K. K.; Goodrich, L.; Lehnert, N. *J. Am. Chem. Soc.* **2009**, 131, 17116. (d) Lehnert, N.; Galinato, M. G.; Paulat, F.; Richter-Addo, G. B.; Sturhahn, W.; Xu, N.; Zhao, J. *Inorg. Chem.* **2010**, 49, 4133.

(11) Oxgaard, J.; Wiest, O. *J. Phys. Chem. A* **2001**, 105, 8236.

(12) (a) Scheidt, W. R.; Barabanschikov, A.; Pavlik, J. W.; Silvernail, N. J.; Sage, J. T. *Inorg. Chem.* **2010**, 49, 6240. (b) Pavlik, J. W.; Barabanschikov, A.; Oliver, A. G.; Alp, E. E.; Sturhahn, W.; Zhao, J.; Sage, J. T.; Scheidt, W. R. *Angew. Chem., Int. Ed.* **2010**, 49, 4400.

(13) Ghosh, A. *J. Biol. Inorg. Chem.* **2006**, 11, 712.

(14) (a) Ellison, M. K.; Scheidt, W. R. *J. Am. Chem. Soc.* **1997**, 119, 7404. (b) Scheidt, W. R.; Duval, H. F.; Neal, T. J.; Ellison, M. K. *J. Am. Chem. Soc.* **2000**, 122, 4651.

(15) (a) Patchkovskii, S.; Ziegler, T. *Inorg. Chem.* **2000**, 39, 5354. (b) Silvernail, N. J.; Barabanschikov, A.; Sage, J. T.; Noll, B. C.; Scheidt, W. R. *J. Am. Chem. Soc.* **2009**, 131, 2131.

(16) (a) Silvernail, N. J.; Pavlik, J. W.; Noll, B. C.; Schulz, C. E.; Scheidt, W. R. *Inorg. Chem.* **2008**, 47, 912. (b) Silvernail, N. J.; Olmstead, M. M.; Noll, B. C.; Scheidt, W. R. *Inorg. Chem.* **2009**, 48, 971.

(17) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, Revision A.02; Gaussian, Inc.: Wallingford, CT, 2009.

(18) Achterhold, K.; Keppler, C.; Ostermann, A.; Burck, U.; Sturhahn, W.; Alp, E. E.; Parak, F. G. *Phys. Rev.* **2002**, E65, 051916.

(19) Merrick, J. P.; Moran, D.; Radom, L. *J. Phys. Chem. A* **2007**, 111, 11683.

(20) (a) Becke, A. D. *Phys. Rev.* **1988**, A38, 3098. (b) Perdew, J. P. *Phys. Rev. B* **1986**, 33, 8822.

(21) (a) Adamo, C.; Barone, V. *J. Chem. Phys.* **1998**, 108, 664. (b) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. *Phys. Rev. B* **1992**, 46, 6671.

(22) (a) Becke, A. D. *Phys. Rev.* **1988**, A38, 3098. (b) Becke, A. D. *J. Chem. Phys.* **1993**, 98, 1372. (c) Becke, A. D. *J. Chem. Phys.* **1993**, 98, 5648. (d) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev.* **1988**, B37, 785.

(23) (a) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, 77, 3865. (b) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1997**, 78, 1396.

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

(25) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, 125, 194101: 1.

(26) Iikura, H.; Tsuneda, T.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2001**, 115, 3540.

(27) Chai, J.-D.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2008**, 10, 6615.

(28) The basis sets can be found in the following references: (a) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, 82, 270–283. *ibid.*, 82, 299. (b) Wadt, W. R.; Hay, P. J. *J. Chem. Phys.* **1985**, 82, 284. (c) Schafer, A.; Horn, H.; Ahlrichs, R. *J. Phys. Chem.* **1992**, 97, 2571. (d) Schafer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, 100, 5829.

(29) Sage, J. T.; Paxson, C.; Wyllie, G. R. A.; Sturhahn, W.; Durbin, S. M.; Champion, P. M.; Alp, E. E.; Scheidt, W. R. *J. Phys.: Condens. Matter* **2001**, 13, 7707.

(30) Li, J. F.; Peng, Q.; Barabanschikov, A.; Pavlik, J. W.; Alp, E. E.; Sturhahn, W.; Zhao, J.; Schulz, C. E.; Sage, J. T.; Scheidt, W. R. *Chem.—Eur. J.* **2011**, 17, 11178.

(31) Jacquemin, D.; Prepete, E. A.; Ciofini, I.; Adamo, C.; Valero, R.; Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2010**, 6, 2071.

Infrared Spectroscopy of Fluxional Molecules from (ab Initio) Molecular Dynamics: Resolving Large-Amplitude Motion, Multiple Conformations, and Permutational Symmetries

Gerald Mathias,^{*,†,‡} Sergei D. Ivanov,^{†,§} Alexander Witt,^{†,||} Marcel D. Baer,^{†,⊥} and Dominik Marx[†]

[†]Lehrstuhl für Theoretische Chemie, Ruhr-Universität Bochum, 44780 Bochum, Germany

ABSTRACT: The computation of vibrational spectra of complex molecules from time correlation functions generated by ab initio molecular dynamics simulations has made lively progress in recent years. However, the analysis of such spectra, i.e., the assignment of vibrational bands to atomic motions, is by no means straightforward. In a recent article [*J. Chem. Theory Comput.* **2011**, *7*, 2028–2039], Mathias and Baer presented a corresponding analysis method that derives generalized normal coordinates (GNCs) from molecular dynamics trajectories, which furnish band positions, band shapes, and infrared intensities of the separated vibrational modes. This vibrational analysis technique relies on the usual quasi-rigidity assumption; i.e., atomic motions are described by small oscillations around a single reference structure. This assumption, however, breaks down if the molecule undergoes large-amplitude motion and visits different conformations along the trajectory or if the same conformation can be adopted by a different ordering of the atoms, i.e., if permutational symmetries have to be considered. Here, we present an extension of the GNC method that handles such cases by considering multiple reference structures, both for different conformations and for permutational symmetries. By introducing a projection technique and computing probabilities that assign the time frames of the trajectories to these reference structures, the vibrational spectra are split into conformational contributions via a consistent time correlation formalism. For each conformation, the permutational symmetries are resolved, which permits one to determine conformation-local GNCs for the band assignment. The working principle and the virtues of this generalization are demonstrated for the simple case of a methyl group rotation. This is followed by an application to a more intricate case: Upon replacing one proton by a deuteron in protonated methane, CH₅⁺, significant changes of its infrared spectrum have been observed since the CH₄D⁺ isotopologue features five different isotopomers. Here, a total of 120 conformational and permutational references are required in the projection scheme in order to capture the frequent and versatile structural transitions of this small but utmost floppy molecule and to assign its infrared spectrum. The extended GNC method is general. Thus, it can be applied readily to systems that require more than one reference structure, and it can be transferred to other theoretical spectroscopies that are formulated in terms of time correlation functions.

1. INTRODUCTION

Vibrational spectroscopy is a powerful experimental technique to monitor and explore (bio)chemical processes and molecular structures.^{1–6} The computation and especially the decoding of such spectra is a challenge to theory, in particular if the underlying potential energy surfaces are strongly anharmonic, permit large-amplitude motion, or are significantly modified by interactions within (micro- or bulk) solvent environments. In recent years, ab initio molecular dynamics (AIMD) simulations⁷ have emerged as a promising approach to theoretical vibrational spectroscopy, especially for systems that are too complex or contain too many coupled degrees of freedom^{8–28} to be readily assessed by quasi-exact quantum mechanical treatments of nuclear motion. The appeal of the AIMD approach is that it works almost out of the box after choosing an appropriate electronic structure method. Since energies, forces, and dipoles can be computed on the fly, AIMD does not require the parametrization of a potential energy surface (PES) and dipole surfaces, which is commonly a challenging^{29–31} and sometimes unfeasible task, if the number of atoms grossly exceeds the current state-of-the-art capabilities of PES parametrization, e.g., in the condensed phase. Moreover, even fairly small but fluxional (“floppy”) systems can be truly challenging in this respect due to the presence of multiple similarly important minima or broad and shallow regions on the PES, which all require special care in the course of parametrization.³²

Of course, the AIMD approach to theoretical vibrational spectroscopy also faces drawbacks because of its classical approximation of quantum-mechanical motion of the nuclei, while obeying quantum mechanics for the electrons. As a result, the classical Maxwell–Boltzmann distribution may grossly differ from the proper quantum-mechanical Bose–Einstein distribution in the relevant parts of the frequency spectrum, not to mention neglecting symmetry effects on rotations and ro-vibrations due to nuclear spin. On the one hand, due to this discrepancy, the vibrational bands of strongly anharmonic systems may exhibit a pronounced temperature dependence in classical simulations,^{33–35} which has to be carefully checked. This cross-checking, on the other hand, may provide deeper insights into the mutual anharmonic couplings of vibrational modes, and thus it may even serve as a sort of analysis tool.³⁵

Although the computation of infrared (IR)^{36,37} and Raman^{36,38} cross-sections from molecular dynamics is both well-defined and straightforward in the frame of the “Heisenberg approach to theoretical spectroscopy”, the assignment of the resulting bands to atomic motions is not. In recent years, a few techniques have been suggested to derive approximate normal modes from molecular dynamics simulations and thereby to assign the vibrational

Received: September 21, 2011

Published: November 29, 2011

bands in the spectra.^{16–18,21,39–43} These methods all rely on an equipartition assumption to a lesser or greater extent, which leads to problems in cases when equipartition is hard to achieve.⁴¹ In the preceding paper⁴⁴ (paper I), Mathias and Baer introduced an approach that is free from such assumptions. The method constructs generalized normal coordinates (GNCs) for the vibrational analysis by an orthonormal transform of mass-weighted coordinates. The transformation is constructed such that it minimizes the norm of the GNCs mutual Fourier-transformed cross-correlation functions. The prospects of the GNC algorithm had been exemplified by the vibrational analysis of *trans*-isoprene (2-methyl-1,3-butadiene), where the IR bands computed from AIMD had been assigned to the underlying atomic motions. In addition, it has already been successfully applied to assign the IR spectra of floppy molecules and complexes such as protonated methane as well as microsolvated hydronium and Zundel cations.^{24,27,35}

Importantly, the GNC algorithm presented in paper I, as well as all other predecessor techniques, is based on the commonly used assumption of quasi-rigidity, namely, that it is sufficient to consider a single reference structure of the molecule and that the dynamical motion can be expressed as (small-amplitude) oscillations around this reference. This assumption is however not justified if large-amplitude motion overrides small-amplitude oscillations such that the trajectories ultimately visit different conformations of a molecule or intramolecular chemical rearrangements lead to new configurations of the atoms. All of these changes correspond to switching between *different minima* on the PES. Examples for conformational changes are *cis*–*trans* isomerizations of polyenes or peptides, whereas different configurations (isomers) may result from intramolecular proton transfer or other chemical reactions. Note that, only for simplicity, we will use the term “conformations” also for different configurations in the aforementioned sense.

A similar problem is rooted in the exploration of the PES by classical dynamics as such. Here, the trajectory can visit chemically *equivalent minima* of the PES, which are related by a permutation of the numbering of equivalent atoms. This problem was already faced in the vibrational analysis of the seemingly simple molecule *trans*-isoprene in paper I. Because of the low rotational barrier of its methyl group (about 11 kJ mol^{−1}),⁴⁵ the latter rotates nearly freely under ambient conditions and thus visits three chemically equivalent minima of the PES (degenerate equilibrium structures) separated by methyl group rotations of 120°. The assumption of quasi-rigidity employed in paper I did not account for these different minima, and correspondingly, the methyl group vibrations were not properly disentangled. As a result of this approximation, they still showed correlations among the corresponding stretching and bending modes. Employing internal coordinates for the vibrational analysis in paper I separated the symmetric modes, e.g., the symmetric methyl stretch, but the antisymmetric modes were still coupled.⁴⁴

For the vibrational analysis of the aforementioned cases, one requires not only one but distinct reference structures for each conformation. If multiple reference structures would be considered, the dynamics along the trajectory could again be described as small-amplitude oscillations around the respective reference structure during the residence time within the corresponding “conformation”, complemented by the transitions between all of the “conformations” available. Using the same idea, one would be able to handle permutational symmetries, which would be represented by additional (permuted) reference structures for each

and every conformation. If the total vibrational spectrum would be dissected into separate contributions of each conformation in this manner, it could be analyzed in terms of the ensemble of thermally populated conformations.

Of course, simply chopping the trajectories into pieces corresponding to a specific instantaneous conformation would be sufficient only if the residence times in each conformation were long compared to the typical periods of the vibrational modes. In this case, the total spectrum could be synthesized a posteriori, including its band assignment, by simply adding the (assigned) conformation-specific spectra using weights according to their free energy differences as obtained from their relative contribution to the total trajectory. However, well-known intricacies in the assignment of IR spectra arise in the opposite limit where switching between conformations is rather frequent on the time scale set by small-amplitude oscillatory motion. In a sense, this limit even defines the aforementioned class of floppy or fluxional molecules that are characterized by effects due to large-amplitude motion. For such cases, a more elaborate and general approach is clearly required.

In this article, we present a generalization of the GNC scheme that can handle systematically multiple reference structures in the vibrational analysis. These structures can either define different conformations of the molecule or represent permutations of atoms of the same species within a given conformation. The extended GNC method is useful for the plethora of cases where a single equilibrium or reference structure is not sufficient to understand vibrational spectra. Furthermore, the very same idea can be readily transferred to other analysis schemes of response properties of molecular systems which are based on time correlation functions as required by the Heisenberg approach to theoretical spectroscopy.

In the following theory section, we first discuss how to dissect the computed IR absorption spectrum into conformational contributions. Then, a similar formalism is developed to compute auto- and cross-correlation functions of the mass-weighted velocities that are specific for each conformation and each permutational pattern. Recombining the permutation-specific correlation functions by employing proper symmetry operations yields correlation functions that are specific only to the conformations. These correlation functions are then used to determine the GNCs of the respective conformations. Further, a projection scheme is introduced that determines the probability for each frame of the trajectory to belong to a certain conformation and permutational pattern, which is used to split the total spectrum quantitatively into conformation-specific contributions.

In the results part, we show how to resolve the permutational symmetry of the methyl group of isoprene to provide a first simple example and test case of the extended GNC method. A much more challenging application in this respect is the analysis of the IR spectra of the small but highly fluxional protonated methane molecule, CH₅⁺,^{17,46–53} and its partially deuterated isotopologues,⁵⁴ CH_nD_{5−n}⁺, *n* = 0–5, for which this generalization was first successfully employed.²⁴

Because of the low intramolecular barriers,⁵⁵ CH₅⁺ undergoes vivid scrambling dynamics; that is, all protons visit each of the five possible binding sites during the simulation. This leads to 5! = 120 reference structures that are necessary for the vibrational analysis of this small molecule. The merits of the extended GNC approach are exemplified on the partially deuterated CH₄D⁺ isotopologue, leading to the fully assigned IR spectrum. To achieve this, the total IR spectrum is split into the contributions

of its isotopomers,⁵⁴ for which in turn the vibrational bands are assigned separately.

2. THEORY

We start the derivation of the multiple configuration (extended) GNC scheme by first developing a formalism to split the computed IR spectrum into the underlying conformational contributions. The linear total IR absorption coefficient³⁶

$$\alpha^{\text{QM}}(\omega) = \frac{2\pi\omega(1 - e^{-\hbar\beta\omega})}{3V\hbar c n(\omega)} \int_{-\infty}^{\infty} dt e^{-i\omega t} \langle \hat{\mathbf{M}}(0) \hat{\mathbf{M}}(t) \rangle \quad (1)$$

is given within the Heisenberg approach to theoretical spectroscopy by the Fourier transform of the autocorrelation function of the time-dependent Heisenberg dipole operator $\hat{\mathbf{M}}(t)$, where $\langle \cdot \rangle$ denotes the quantum-statistical ensemble average at temperature T . The prefactor depends on the sample volume V , the refractive index $n(\omega)$, and the inverse temperature $\beta = 1/k_B T$. To obtain a corresponding $\alpha(\omega)$ from classical dynamics trajectories, one approximates the autocorrelation function of Heisenberg's dipole operator $\hat{\mathbf{M}}(t)$ by the autocorrelation function of the dipole moment $\boldsymbol{\mu}(t)$. Here, this replacement is done after rewriting eq 1 in terms of the Kubo-transformed quantum correlation function, which readily yields

$$\alpha(\omega) = \frac{2\pi\beta\omega^2}{3Vcn(\omega)} \int dt e^{-i\omega t} \langle \boldsymbol{\mu}(0) \boldsymbol{\mu}(t) \rangle \quad (2)$$

as our basic working equation.⁵⁶ Note that many different so-called “quantum corrections factors” have been introduced in the literature mainly to impose the detailed balance condition and that the “harmonic quantum correction” corresponds to the prefactor in eq 2, see ref 56.

Our first goal is to determine individual absorption coefficients $\alpha_{\xi}(\omega)$ for each reference conformation ξ . Thus, we have to split the dipole autocorrelation function into conformational contributions. Given that we can classify the frames of a trajectory, and thereby assign each frame at time t with a certain probability $p_{\xi}(t)$ to conformation ξ satisfying

$$\sum_{\xi} p_{\xi}(t) = 1, \quad 0 \leq p_{\xi}(t) \leq 1 \quad (3)$$

we can split the ensemble average of the dipole autocorrelation function

$$\langle \boldsymbol{\mu}(0) \boldsymbol{\mu}(t) \rangle = \sum_{\xi} \langle p_{\xi}(0) \boldsymbol{\mu}(0) \boldsymbol{\mu}(t) \rangle \quad (4)$$

into a sum of conformation-specific correlation functions. These specific correlation functions are restricted to trajectories visiting conformation ξ at $t = 0$. Accordingly, we define the conformation specific IR absorption coefficient as

$$\alpha_{\xi}(\omega) = \frac{2\pi\beta\omega^2}{3Vcn(\omega)} \int dt e^{-i\omega t} \langle \boldsymbol{\mu}_{\xi}(0) \boldsymbol{\mu}(t) \rangle \quad (5)$$

where we have introduced the probability-weighted dipole moment $\boldsymbol{\mu}_{\xi}(t) \equiv p_{\xi}(t) \boldsymbol{\mu}(t)$. Thus, if one can classify the trajectories according to conformations, the computation of conformation-specific spectra is straightforward and mathematically exact. We will discuss possible projection strategies further below. Note that the Fourier transform of the correlation functions introduced here and in the following can be efficiently calculated

by employing the cross-correlation theorem as discussed in detail in paper I.

For the subsequent vibrational analysis within the extended GNC scheme, which is the core task for any spectral assignment, the splitting procedure becomes slightly more complicated. We recall from paper I that for GNC we have to compute a tensorial version of the vibrational density of states (VDOS) $\Theta: \mathbb{R} \rightarrow \mathbb{R}^{3N} \times \mathbb{R}^{3N}$ for N atoms given by

$$\Theta(\omega) = \frac{\beta}{\pi} \int dt e^{-i\omega t} \langle \dot{\mathbf{c}}(0) \otimes \dot{\mathbf{c}}(t) \rangle \quad (6)$$

with $\dot{\mathbf{c}}(t) = \mathbf{M}^{1/2} \mathbf{v}(t)$, $M_{ij} = \delta_{ij} m_j$, being the trajectory of the mass-weighted atomic velocities \mathbf{v} . Here, the correlations are evaluated as an outer product $\dot{\mathbf{c}} \otimes \dot{\mathbf{c}} = \dot{\mathbf{c}} \dot{\mathbf{c}}^T$ of the mass-weighted velocity components, yielding their autocorrelations on the diagonal of Θ and their mutual cross-correlations in the off-diagonal elements of Θ . To obtain the generalized normal coordinates, Θ is brought to diagonal form as close as possible by an orthonormal transform of the $\dot{\mathbf{c}}$, thereby minimizing the cross-correlations.⁴⁴

For a conformation-specific VDOS, i.e., for $\Theta_{\xi}(\omega)$, we would like to employ the same splitting procedure as we have done to obtain the corresponding IR spectra, i.e., α_{ξ} . However, we have to additionally consider possible permutational symmetries, i.e., reference structures $\mathbf{x}_{\xi,i}^{\text{ref}}$ belonging to the same conformation ξ but differing from a default reference $\mathbf{x}_{\xi,0}^{\text{ref}}$ by a permutation

$$\mathbf{x}_{\xi,i}^{\text{ref}} = \mathcal{P}_{\xi,i} \mathbf{x}_{\xi,0}^{\text{ref}} \quad (7)$$

where $\mathcal{P}_{\xi,i}$ is a permutation operator interchanging only the positions of atoms of the same species. Therefore, we have to classify the trajectory additionally according to the permutational pattern i with probabilities $p_{\xi,i}$ which satisfy

$$\sum_i p_{\xi,i}(t) = p_{\xi}(t) \quad (8)$$

We can weight the $\dot{\mathbf{c}}$ with these probabilities, which gives $\dot{\mathbf{c}}_{\xi,i}(t) = p_{\xi,i}(t) \dot{\mathbf{c}}(t)$, and introduce conformation- and permutation-specific

$$\Theta_{\xi,i}(\omega) = \frac{\beta}{\pi} \int dt e^{-i\omega t} \langle \dot{\mathbf{c}}_{\xi,i}(0) \otimes \dot{\mathbf{c}}(t) \rangle \quad (9)$$

which sum up to the global Θ .

Having to evaluate a separate $\Theta_{\xi,i}(\omega)$ for each permutational pattern is however neither desirable nor necessary. Employing the back-transforms $\mathcal{P}_{\xi,i}^{-1}$, we can bring all permutational patterns to the same reference structure and calculate

$$\Theta_{\xi}(\omega) = \sum_i \mathcal{P}_{\xi,i}^{-1} \Theta_{\xi,i}(\omega) \mathcal{P}_{\xi,i} \quad (10)$$

as the sum of the transformed $\Theta_{\xi,i}$ or computationally more convenient

$$\Theta_{\xi}(\omega) = \frac{\beta}{\pi} \int dt e^{-i\omega t} \sum_i \langle \mathcal{P}_{\xi,i}^{-1} \dot{\mathbf{c}}_{\xi,i}(0) \otimes \mathcal{P}_{\xi,i}^{-1} \dot{\mathbf{c}}(t) \rangle \quad (11)$$

Sometimes it is preferential to analyze the molecular motion in terms of a set of internal coordinates $\mathbf{s}_{\xi}(\mathbf{x}) = (s_{\xi,1}(\mathbf{x}), \dots, s_{\xi,3N-6}(\mathbf{x}))$, specifically chosen for conformation ξ (see paper I for details). Here, the permutational symmetries can be resolved in two ways: either one applies the permutation operators directly to the Cartesians

$$\mathbf{s}_{\xi,i}(\mathbf{x}) = \mathbf{s}_{\xi}(\mathcal{P}_{\xi,i}^{-1} \mathbf{x}) \quad (12)$$

and then evaluates the internal coordinates, which is simple to implement but may involve many redundant evaluations of internal coordinates, or alternatively, $\mathcal{P}_{\xi,i}$ can directly act on the definitions of the \mathbf{s}_{ξ} to give

$$\mathbf{s}_{\xi,i}(\mathbf{x}) = (\mathcal{P}_{\xi,i}\mathbf{s}_{\xi})(\mathbf{x}) \quad (13)$$

That is, one transforms the internal coordinate definitions similar to what has been done with the reference structure according to eq 7. Such internal coordinates are commonly linear combinations of primitive coordinates such as bond lengths, angles, and dihedrals, where a large part of these primitives will be invariant under coordinate permutation or the permuted primitive will be already present in the set of primitives corresponding to the reference structure. Only few primitives will lead to new primitives upon employing the permutation operator. Thus, the set of all relevant primitives has to be determined once for the whole trajectory, and the internal coordinates for a given permutational reference i can be efficiently calculated from these primitives. On the basis of these internal coordinates, the tensorial VDOS evaluates to

$$\Theta_{\xi}(\omega) = \frac{\beta}{\pi} \int dt e^{-i\omega t} \sum_i \langle p_{\xi,i}(0) \mathbf{K}^+ \mathbf{s}_{\xi,i}(0) \otimes \mathbf{K}^+ \mathbf{s}_{\xi,i}(t) \rangle \quad (14)$$

where \mathbf{K}^+ with $\mathbf{G}^{-1} = (\mathbf{K}^+)^T \mathbf{K}^+$ is a root of the inverse Wilson \mathbf{G} matrix that transforms the internal coordinates back to Cartesians.^{44,57} Note here that we have employed the common approximation⁵⁷ $\mathbf{G} = \mathbf{G}[\mathbf{s}_{\xi,0}] \approx \mathbf{G}[\mathbf{s}_{\xi,0}(\mathbf{x} = \mathbf{x}_{\xi,0}^{\text{ref}})]$; i.e., we evaluate \mathbf{G} solely at the position of the reference structure $\mathbf{x}_{\xi,0}^{\text{ref}}$ and neglect the dependence of this matrix on the internal coordinates.

What remains is to introduce a computationally practical projection scheme to compute the probabilities $p_{\xi}(t)$ and $p_{\xi,i}(t)$. Here, we choose a general form

$$p_{\xi}(t) = \frac{\exp\left(-\frac{1}{2\sigma_c^2} \min_i d_c(\mathbf{x}(t), \mathbf{x}_{\xi,i}^{\text{ref}})^2\right)}{\sum_{\eta} \exp\left(-\frac{1}{2\sigma_c^2} \min_j d_c(\mathbf{x}(t), \mathbf{x}_{\eta,j}^{\text{ref}})^2\right)} \quad (15)$$

where $d_c(\mathbf{x}, \mathbf{x}')$ is a metric measuring the distance between \mathbf{x} and \mathbf{x}' and the sum over η in the denominator runs over all conformations considered. In order to determine the distance of a given frame to a conformation, we have to keep in mind its permutational symmetries. Therefore, we take the minimum of the distances to all possible permutational references i of a given conformation ξ . Note that $p_{\xi}(t)$ is properly normalized according to eq 3 by the sum of Gaussians in the denominator.

The width of the Gaussians σ_c determines the width of the switching region between two neighboring conformations. Correspondingly, large values of σ_c lead to slow and smooth transitions, whereas small values of σ_c lead to fast but possibly noisy transitions if the system oscillates in the switching region between two conformations. For the permutational probabilities, we choose similarly

$$p_{\xi,i}(t) = p_{\xi}(t) \frac{\exp\left(-\frac{1}{2\sigma_p^2} d_p(\mathbf{x}(t), \mathbf{x}_{\xi,i}^{\text{ref}})^2\right)}{\sum_{j=1}^{j_{\max}} \exp\left(-\frac{1}{2\sigma_p^2} d_p(\mathbf{x}(t), \mathbf{x}_{\xi,j}^{\text{ref}})^2\right)} \quad (16)$$

Finally, proper metrics $d_c(\cdot, \cdot)$ and $d_p(\cdot, \cdot)$ between the coordinate frames and the reference coordinates have to be chosen. These metrics should return similar nearest-neighbor distances

$$r_{\xi}^{\text{nn}} = \min_{\eta,j} d_c(\mathbf{x}_{\eta,j}^{\text{ref}}, \mathbf{x}_{\xi,0}^{\text{ref}}) \quad (17)$$

for all conformations and

$$r_{\xi,i}^{\text{nn}} = \min_j d_p(\mathbf{x}_{\xi,j}^{\text{ref}}, \mathbf{x}_{\xi,i}^{\text{ref}}) \quad (18)$$

for all permutations of a given reference, such that transitions between two reference structures occur on similar length scales. These length scales can then be used to choose σ_c and σ_p .

One obvious choice for d_c or d_p is the mass-weighted root-mean-square deviation

$$d_{\text{RMS}}(\mathbf{x}, \mathbf{x}') = \min_{\mathbf{A} \in \mathcal{O}_3, \mathbf{t} \in \mathbb{R}^3} \sqrt{\frac{\sum_{k=1}^N m_k (\mathbf{A}\mathbf{r}_k + \mathbf{t} - \mathbf{r}'_k)^2}{\sum_{k=1}^N m_k}} \quad (19)$$

where \mathbf{t} is a translation and \mathbf{A} is an orthonormal transform acting on the atomic coordinates \mathbf{r}_k of \mathbf{x} , thereby aligning it with \mathbf{x}' . Here, all internal degrees of freedom of the molecule contribute to the distance between structures \mathbf{x} and \mathbf{x}' . Of course, mass weighting can be omitted or replaced by another weighting scheme if this seems more suitable for the specific problem. Alternatively, one can choose a metric

$$d_{\text{IC}}(\mathbf{x}, \mathbf{x}') = \sqrt{\sum_{k=1}^{N_{\text{IC}}} a_k (s_k(\mathbf{x}) - s_k(\mathbf{x}'))^2} \quad (20)$$

based on a set of N_{IC} internal coordinates s_k . The a_k 's provide the possibility to combine different kinds of internal coordinates, e.g., bonds and dihedrals, by weighting their contributions. The evaluation of the internal coordinates does not require a structural alignment, and in many cases the choice of the s_k is straightforward. For example, to determine the rotation of a methyl group, one would choose the corresponding torsional angles connecting the protons and the carbon to the rest of the molecule. Similarly simple internal coordinates are also found for many typical conformational changes such as *cis/trans* isomerizations in polyenes and peptides or transitions between ring puckers of sugars.⁵⁸

3. COMPUTATIONAL METHODS

As outlined in the Introduction, we consider two problems, which are challenging to our algorithm due to fast transitions between reference structures, either due to switching between permutational patterns or different conformations or a combination of both.

The first illustrative example is the complete vibrational analysis of the methyl group of isoprene. Here, we use the data set described in paper I, comprising 41 trajectories of isoprene in the gas phase, each spanning 25 ps of Born–Oppenheimer ab initio dynamics.⁷ Initial conditions have been drawn from the canonical ensemble at 300 K. The electronic structure was treated by the BLYP density functional at a density cutoff of 280 Ry within the CP2k simulation package;⁵⁹ details are given in paper I.

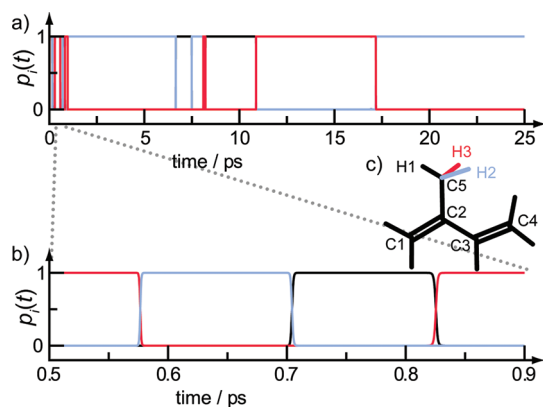


Figure 1. Probabilities $p_i(t)$ for the three permutational references of the methyl group of isoprene for one representative trajectory. In the default reference (black curve), whose structure is shown in panel c, proton H1 lies in the symmetry plane of the molecule. The other two references are obtained by rotating the methyl group by 120° , such that either H2 (blue curve) or H3 (red curve) is in the symmetry plane. The upper panel a shows the full trajectory, whereas b zooms in on a part of the first picosecond, when the methyl group rotates fast.

The second, more challenging example is the vibrational analysis of the CH_4D^+ isotopologue of protonated methane. The isolated CH_4D^+ molecule was simulated by Born–Oppenheimer MD as implemented in the CPMD⁶⁰ package. Here, the well-established setup discussed in detail in ref 17 was employed, i.e., the Perdew–Zunger local density approximation supplemented by Becke’s exchange gradient correction at a plane–wave cutoff of 35 Ry, as used in previous publications on protonated methane.^{15,24,28,46} Starting from canonically distributed initial conditions at a temperature of $T = 110$ K, 297 trajectories have been integrated for 10.2 ps with a time step of $\Delta t = 0.34$ fs.

Both systems have been analyzed with the program normcor, which was extended by the algorithms described in this article. Important parameters that have been chosen for these vibrational analyses will be discussed along with the results.

4. RESULTS AND DISCUSSION

To illustrate the ingredients of the extended GNC algorithm, we will lead through the vibrational analysis of the two examples that address the aspects of permutational symmetries and permutational symmetries combined with conformational changes.

4.1. Methyl Group Rotation: *trans*-Isoprene. We start the discussion with the comparatively simple case of a methyl group rotation, which switches only between chemically equivalent minima, and one does not have to consider different conformations. Already in paper I, it was shown that the vibrational modes of a methyl group cannot be clearly resolved if only one reference structure is considered. Because of the low barrier of the internal rotation of the methyl group, which is about 11 kJ mol^{-1} experimentally^{45,61} and about 10 kJ mol^{-1} in our calculation employing the BLYP functional, the methyl group can easily cross this barrier under ambient conditions. Figure 1c introduces the atom labeling scheme and illustrates that due to the C_s symmetry of isoprene the three methyl protons cannot be considered equal. The two protons outside the mirror plane, H2 and H3, are at chemically equivalent sites but differ from the in-plane proton, H1. The rotation of the methyl group by 120° will therefore exchange the character of the three protons, which

mathematically represents a cyclic permutation of their numbering. Correspondingly, two additional reference structures have to be considered for the methyl group, which yield the protons H2 and H3, respectively, at the in-plane position occupied by H1 in the default reference.

To discriminate between the reference structures along the trajectory, we used the metric (c.f. eq 20)

$$d_{\text{IC}}(\mathbf{x}, \mathbf{x}') = \sqrt{\sum_{j=1,3} \sum_{i=1,2,3} (\gamma_{\text{H}_i, \text{C}_j}(\mathbf{x}) - \gamma_{\text{H}_i, \text{C}_j}(\mathbf{x}'))^2} \quad (21)$$

constructed from all six dihedral angles $\gamma_{\text{H}_i, \text{C}_j}$ across the C5–C2 bond. The permutation probabilities $p_{\xi,i}(t)$ according to eq 16 were computed using a width $\sigma_p = 15^\circ$ of the Gaussians; note that $p_{\xi}(t) = 1$ because only one conformation of isoprene is considered.

Figure 1 illustrates the resulting probabilities for a selected trajectory. The whole trajectory in Figure 1a is characterized by long periods where only one reference structure is present. However, during the first picosecond, enlarged in Figure 1b, the methyl group rotates fast, and in turn, protons H3 (red curve) followed by H2 (blue), H1 (black), and finally again H3 (red) are located in the mirror plane. The transitions between the reference structures occur on a time scale of only 10 fs, during which time the probabilities switch continuously and monotonously between zero and one. During this first picosecond of this particular trajectory, the methyl group contains enough energy in its rotational degree of freedom to cross the barrier between two adjacent reference structures. Since the barriers between all reference structures have the same height, the methyl group visits all three reference structures in turn. When the methyl group loses rotational energy due to couplings to other modes, its energy does not suffice any longer for barrier crossings, and thus the molecule gets transiently trapped in one conformation, e.g., during the interval from 2 to 7 ps in Figure 1a, until the methyl group again gains enough energy from other modes to cross the barrier. This behavior, i.e., periods with frequent switching between the reference structures followed by longer residence times within one reference structure, is also found in the other 40 trajectories. Thus, sufficient sampling of initial conditions from a well thermalized trajectory needs to be carried out in order to properly compute canonically averaged infrared spectra.

After computing the tensorial VDOS from eq 11, which combines the contributions of the three reference structures, and minimizing its off-diagonal norm (see paper I), one obtains the bands of the generalized normal modes on the diagonal of $\Theta(\omega)$. Figure 2 compares the nine vibrational modes associated with the methyl group obtained by (a) using a single reference structure (i.e., plain GNC according to paper I) and (b) considering three reference structures according to the extended GNC method introduced here. For both cases, we have employed a RMSD fitting procedure to remove the global translation and rotation of the molecule; similar results are obtained by using an internal coordinate transform instead (not shown).⁴⁴ Already the analysis with a single reference structure in Figure 2a, i.e., using plain GNC, shows the dominant characteristics of the modes. However, their spectra are not singly peaked but show sidebands at the position of other modes, which hint at a remaining coupling that cannot be disentangled by the plain GNC analysis. For example, the symmetric stretch (black curve)

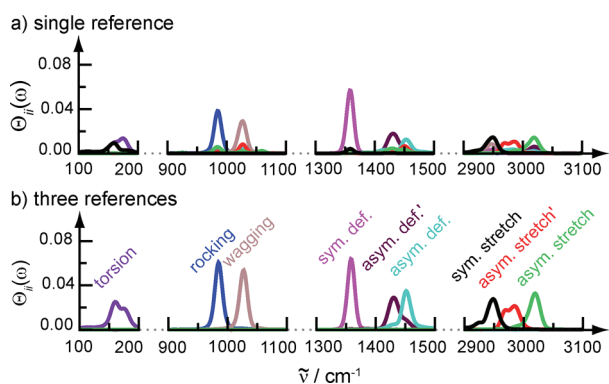


Figure 2. Localized modes $\Theta_{ii}(\omega)$ obtained for (a) a single reference (as introduced in paper I) and (b) for three references within the extended GNC algorithm introduced here. The color code defines the mode labels and is the same for both graphs.

located at 2950 cm^{-1} has a small sideband at 150 cm^{-1} and seems to couple to the methyl torsion, which is also visible in the associated normal coordinate vectors \mathbf{q}_i . To evaluate the quality of the normal coordinates, we have computed their overlaps

$$\vartheta_i = |\mathbf{q}_i \cdot \mathbf{q}_i^h| \quad (22)$$

with the corresponding normal coordinates \mathbf{q}_i^h obtained from the harmonic analysis described in paper I. Since the normal coordinates are normalized, $\vartheta_i = 1.0$ is obtained for $\mathbf{q}_i = \pm \mathbf{q}_i^h$. The average over the nine methyl modes yields only $\langle \vartheta \rangle = 0.25$, implying that the normal coordinates of these modes strongly differ from the harmonic description. Only the symmetric deformation mode shows a significant resemblance at $\vartheta = 0.88$. Interestingly, this mode also shows the best resolved peak in Figure 2a with only very little coupling to other methyl modes.

In stark contrast, employing three reference structures in Figure 2b within the extended GNC scheme completely resolves these couplings, and only a small overlap between the two neighboring asymmetric deformation modes near 1440 cm^{-1} remains. In particular, the normal modes obtained for the deformation and stretching modes reflect the C_s symmetry of the molecule, where we have indicated the modes of odd character by a prime. Note that employing three reference structures hardly changes the peak positions of the vibrational bands, which differ at most by 2 cm^{-1} between Figure 2a and b.

The improved description resulting from using three reference structures also shows up in their overlaps θ_i of the resulting methyl normal coordinates with respect to the harmonic ones. They are larger than 0.98 for eight out of the nine methyl modes, thus indicating a perfect match. Only the methyl torsion around 160 cm^{-1} notably deviates between the extended GNC and the harmonic description with $\vartheta = 0.81$, which is caused by a different coupling to the torsion around the C2–C3 single bond having a similar frequency, 151 cm^{-1} .⁴⁴

Thus, the generalization of the GNC algorithm fully resolves the permutational symmetry of the methyl group and cleanly identifies all modes from the AIMD trajectories. Next, we examine the extended GNC method for a much more involved case where, in addition, multiple conformational states have to be considered.

4.2. Multiple Conformations: CH_4D^+ . Protonated methane, CH_5^+ , is infamous for its vigorous scrambling dynamics, that is, it actively travels between all minima on its shallow PES.^{46,47}

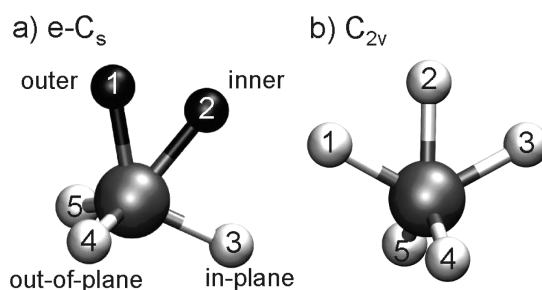


Figure 3. (a) The global minimum structure of the eclipsed C_s (e- C_s) and (b) the C_{2v} transition structure of the CH_5^+ pseudorotation. In the e- C_s structure, the atoms which form the H₂ moiety are marked in black to guide the eye.

The global minimum of the PES is of eclipsed C_s (e- C_s) symmetry, see Figure 3a, and may be visualized as a “CH₃ tripod”, which includes protons H3, H4, and H5 in Figure 3a, to which the so-called “H₂ moiety”, comprising protons H1 and H2 marked in black in Figure 3a, is attached. The H₂ moiety is connected to the tripod via a three-center–two-electron bond, which is significantly weaker than the covalent C–H bonds within the tripod.⁴⁶ The global minimum can be realized via $5! = 120$ permutations of the protons, which all differ from the standpoint of classical dynamics where each atom carries a unique label or number. Most importantly, these 120 minima are connected via extremely low barriers. The transition states between these minima are either a staggered C_s (s- C_s) structure, that is obtained by a rotation of the H₂ moiety by 30° in Figure 3a, or a C_{2v} structure shown in Figure 3b. The respective barrier heights are $\sim 0.1\text{ kcal mol}^{-1}$ (i.e., about 40 cm^{-1} , 0.004 eV or 50 K) and 0.8 kcal mol^{-1} ($\sim 300\text{ cm}^{-1}$, 0.03 eV or 400 K) higher in energy than the e- C_s global minimum.⁵⁵ The transition over the s- C_s barrier is termed internal (H₂ moiety) rotation, and the transition over the C_{2v} structure is referred to as pseudorotation. From the C_{2v} transition state, a new H₂ moiety can be formed by the atoms at positions H2 and H3. Concomitantly, the atom at position H1 occupies the position H3 in the newly formed tripod. Together with the complementary internal rotation, the pseudorotation permits each atom to travel between all five binding sites, which ultimately leads to what is often called “hydrogen scrambling”.

Isotopic substitution complicates the situation even further. Replacing one proton by a deuterium and, thus, forming the CH_4D^+ isotopologue leads to five different isotopomers⁵⁴ which have to be considered, since the deuterium can as well reside at any of the five binding sites. The four protons can then be distributed in $4! = 24$ ways on the remaining binding sites of each isotopomer. Thus, vibrational analysis of CH_4D^+ requires five conformations with 24 permutations each. In order to set the nomenclature, we classify the tripod atoms with respect to the symmetry plane σ_{H_1} ; namely, H4 and H5 are referred to as “out-of-plane” atoms (or “sites”) whereas H3 is referred to as the “in-plane” atom (site), see Figure 3a. Analogously, H2 in the H₂ moiety is considered to be the “inner” atom, since it is placed between H1 (being the “outer” atom) and the “in-plane” atom H3, all being coplanar with C. As the isotopomers of CH_4D^+ are sufficiently classified by specifying the deuterium position, we employ the same terminology to refer to them, e.g., inner isotopomer is the one with the deuterium in the inner position. Note that since the out-of-plane isotopomers differ only by their handedness, they yield the same spectral contributions.

Table 1. Internal Coordinates s_k and Weights a_k of the Metric d_{IC} as Defined in eq 20 for CH_4D^{+a}

k	s_k	a_k	$s_k(\mathbf{x}^0)$
1	stre(1,2)/Å	1.00	0.98
2	stre(3,4)/Å	1.00	1.79
3	stre(3,5)/Å	1.00	1.79
4	stre(4,5)/Å	1.00	1.89
5	stre(1,5)/Å	1.00	1.72
6	stre(1,4)/Å	1.00	1.72
7	stre(1,3)/Å	1.00	2.05
8	stre(2,3)/Å	1.00	1.44
9	stre(2,4)/Å	1.00	1.95
10	stre(2,5)/Å	1.00	1.95
11	tors(4,C,1,5)/rad	0.33	-2.12 (-121°)
12	tors(3,C,2,1)/rad	0.33	3.14 (180°)
13	tors(3,C,4,1)/rad	0.33	2.30 (132°)
14	tors(3,C,5,1)/rad	0.33	-2.30 (-132°)
15	tors(1,C,2,4)/rad	0.33	1.32 (76°)
16	tors(1,C,2,5)/rad	0.33	-1.32 (-76°)

^aReference values $s_k(\mathbf{x}^0)$ are given for the e- C_s structure with the deuteron at the outer position. The coordinates “stre(i,j)” refer to distances between sites i and j . Correspondingly, “tors(i,j,k,l)” refers to the set of dihedral angles. Note that the internal coordinates become dimensionless through the division by the given units (Å, rad).

The first task for our comprehensive vibrational analysis of the CH_4D^+ isotopologue using the extended GNC method is to split its IR spectrum into the contributions of the different isotopomers. For this purpose, we constructed the reference structures of the isotopologues starting from the e- C_s structure with the deuteron at the outer position. The remaining four reference structures were obtained by pair permutations of the deuteron coordinates with the coordinates of each proton. For the assignment of the trajectory frames to the isotopomers, we employed a metric d_{IC} as defined in eq 20 with the internal coordinates s_k and weights a_k given in Table 1.

Here, the structure is determined by all D–H and H–H distances across the molecule, which span a range from 0.98 Å to 2.05 Å in the minimum e- C_s structure. In addition, dihedral angles ($k = 11–16$) help to determine the topology of the deuteron and protons around the carbon. The a_k for the dihedral angles are chosen smaller than those of stretches since the dihedrals vary over a larger numerical range. With this choice, the contributions of both types of internal coordinates are balanced in the metric.

For a proper assignment of the trajectory frames to the isotopomers, we have to ensure that the reference structures are well separated in the space spanned by the internal coordinates, which define the metric d_{IC} . Figure 4 displays the relative occurrence of the distances between all 120 reference structures, including the 24 permutational references for each isotopologue. Most references are well separated by distances around 4.0, which correspond to multiple pair permutations of the deuteron and protons between the compared reference structures. Only a few reference structures are closer together at distances around 1.0 with a minimum distance of 0.9. These closer distances result from interchanging only the deuteron and one proton. These nearest neighbor distances now permit us to choose a proper value for the width of the Gaussians σ_c , which determines the width of the switching region between conformations in eq 3.

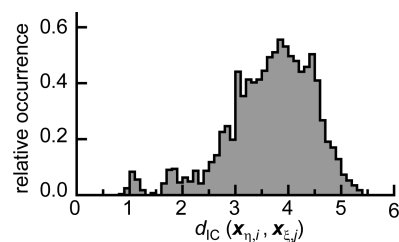


Figure 4. Relative occurrence of the distances between the 120 reference structures. Note that d_{IC} is dimensionless (c.f. Table 1).

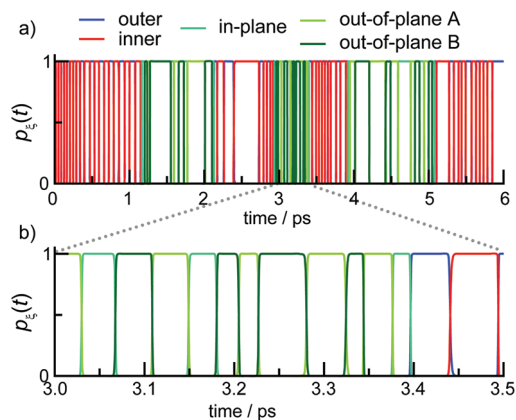


Figure 5. Probabilities $p_\xi(t)$ of the five different isotopomers ξ for one representative trajectory of isotopologue CH_4D^+ akin to Figure 1. The lower panel b zooms in on a selected time interval.

Here, we choose $\sigma_c = 0.1$, which separates nearest neighbors by nine standard deviations of the Gaussians. Note that choosing $\sigma_c = 0.2$ hardly changes the results, and therefore, the algorithm is robust with respect to this parameter. Furthermore, we employed the very same metric to discriminate between the 24 permutational symmetries of each isotopomer, which we will, therefore, not discuss separately.

The resulting probabilities of the isotopomers along the first 6 ps of a representative trajectory are shown in Figure 5a. Due to the employed metric, the probabilities switch fast and continuously between zero and one. As the close-up on the trajectory in Figure 5b shows, the molecule resides within an isotopomer on a time scale of only a few tens of femtoseconds, which is much shorter than the stability of the permutational states of the methyl group in isoprene, c.f. Figure 1. Furthermore, when we follow the sequence of the isotopomers in Figure 5b, we first see a switching between those three isotopomers where the tripod contains the deuteron. These transitions are caused by the internal rotation of the H_2 moiety with respect to the tripod. At about 3.4 ps, we observe a transition from the in-plane isotopomer to the inner isotopomer, which indicates a pseudorotation event that exchanges the deuteron from the tripod with the proton from the H_2 moiety, thus yielding a HD moiety. Thus, the probabilities obtained by our general projection procedure convincingly reflect the scrambling dynamics of the molecule.

In order to quantify its time scales, the time autocorrelation function of the probabilities $p_\xi(t)$ averaged over all five isotopomers ξ of CH_4D^+ is plotted in Figure 6. The autocorrelation function shows a clear biexponential decay behavior with the fitted time constants of about $\tau_1 = 58$ fs and $\tau_2 = 630$ fs. Comparing these time constants with the isotopomer dynamics

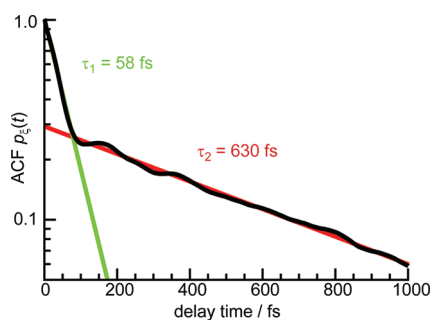


Figure 6. Log-scale time autocorrelation function of the probabilities $p_{\xi}(t)$ averaged over the isotopomers ξ (black curve). Colored lines display two fitted exponentials of time constants $\tau_1 = 58$ fs and $\tau_2 = 630$ fs.

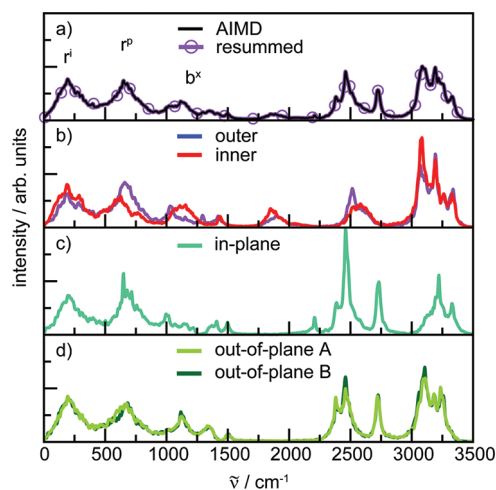


Figure 7. (a) Total IR spectra of CH_4D^+ as generated directly from AIMD simulations at $T = 110$ K (black line) and total spectrum obtained by resumming the five split spectra depicted in panels b–d weighted by the relative occurrences of the isotopomers (magenta circles). The other panels show the five isotopomer spectra obtained by the splitting procedure: the deuteron is contained in a HD moiety in b, whereas it is in the tripod in panels c and d, according to the color code as indicated.

in Figure 5b, one can deduce that τ_1 is associated with the frequent switching between isotopomers that mostly stems from internal rotations. The slower time constant τ_2 is identified by closer inspection of Figure 5a. Here, longer time intervals are dominated by either the inner and outer isotopomers or the in-plane and the two out-of-plane isotopomers. Thus, τ_2 describes the exchange of the deuteron between the tripod and H_2 moiety, which is the pseudorotation of the deuteron. Obviously, the present analysis yields classical lifetimes that are consistent with the classical approximation to the dynamics and the computation of the IR spectrum according to eq 2. It is noted in passing that quantum IR spectra have been computed and quantum lifetimes have been discussed in ref 28.

Having established the computation of the probabilities, we can now split the computed total IR absorption spectrum depicted in Figure 7a into the distinct spectral contributions of all five isotopomers that contribute, which are displayed in Figure 7b–d. Grossly speaking, three spectral regions can be identified in the total IR spectrum depicted in Figure 7a according to ref 24. The region up to roughly 750 cm^{-1} contains so-called rearrangement modes, i.e., the H_2/HD moiety rotation (r^j)

and the pseudorotation motion (r^p), which have been discussed above. The modes between about 750 cm^{-1} and 1500 cm^{-1} can be attributed to bending motions b^x , and the modes above 1500 cm^{-1} are stretching modes of the H_2/HD moiety and the tripod. Particularly the latter are indicative of the different isotopomers because of the different force constants of the weaker bound H_2/HD moiety atoms (due to three-center–two-electron bonding) and the stronger bound tripod atoms, and due to the large frequency shift of the corresponding stretches upon H/D exchange. Thus, already a close inspection of the split spectra should allow for a tentative band assignment. The lowest stretching modes at around 1800 cm^{-1} are found for the outer and inner isotopomers in Figure 7b. Correspondingly, they should be associated with the C–D stretches of the deuteron in the HD moiety. The next modes at 2500 cm^{-1} therein should then represent the C–H stretch of the HD moiety protons, and hence, the modes located above 3000 cm^{-1} belong to the tripod protons. In contrast, the HD moiety C–D stretch at 1800 cm^{-1} is absent in panels c and d, which show the spectra corresponding to the deuterated tripod, but a new band at 2750 cm^{-1} appears. Therefore, one would, at first glance, assign this band to the C–D stretches of the tripod. However, if one considers the reduced mass ratios of a C–H and a C–D bond, a C–H stretch located at around 3200 cm^{-1} would be expected at around 2350 cm^{-1} upon H/D exchange, according to a most simplistic harmonic approximation estimate. Because a blue-shift of 400 cm^{-1} from the estimate seems too large, the band at 2750 cm^{-1} more likely stems from a H_2 moiety stretch. Then, the lower bands below 2500 cm^{-1} should be associated with the remaining C–H stretch of the H_2 moiety and the C–D stretch of the tripod. Going beyond such qualitative discussion by inspection requires one to use the extended GNC analysis scheme, which will provide full and unambiguous assignment of the total IR spectrum in terms of atomic motions as demonstrated below.

Comparing the split spectra of the two out-of-plane isotopomers, which only differ in their handedness and thus should yield strictly identical spectra, one can see that not only is the overall band structure identical, thus supporting the splitting procedure as such, but also that merely a few peak intensities deviate slightly. This implies both that the statistics sampled are sufficient, such that the peak heights of the other isotopomers can also be assumed to be converged, and that the splitting procedure works quantitatively. Note that although the statistics gathered for CH_4D^+ (corresponding to about 3 ns of AIMD trajectory) are 3 times those generated for isoprene, the convergence here is only comparable because the statistics are shared among the five isotopomers.

Finally, Figure 7a also includes the synthetic spectrum obtained by resumming the isotopomer-specific split spectra from panels b to d and weighting them by the relative occurrences of the respective isotopomers in the simulations ranging from 18.6% to 21.4%. These five weights are straightforwardly obtained from expectation values

$$\langle p_{\xi} \rangle = \frac{1}{N_{\text{tr}} t_{\text{max}}} \sum_{k=1}^{N_{\text{tr}}} \int_0^{t_{\text{max}}} dt p_{\xi}^k(t), \text{ for } \xi = 1, \dots, 5 \quad (23)$$

over the $N_{\text{tr}} = 297$ trajectories and satisfy $\sum_{\xi} \langle p_{\xi} \rangle = 1$ by virtue of eq 3. The so-called resummed spectrum matches the total IR spectrum as directly generated by AIMD one to one, which confirms that the projection and splitting procedure introduced here works quantitatively; i.e., IR intensity is neither gained nor lost by the splitting.

Although inspection of the split spectra seems to permit a tentative assignment of certain modes and yields qualitative insights into the structure of the bands as alluded to above, it is still insufficient for the unambiguous assignment even of the stretching modes. Thus, a quantitative assignment based on the GNC analysis is required to fully resolve the structure of these isotopomer spectra, as will be shown in the following. Here, we will limit the discussion of this analysis to the stretching modes, albeit the algorithm equally assigns the low frequency region; see e.g. ref 24 for the discussion of all modes of the CHD_4^+ isotopologue. Similar to the extended GNC analysis of isoprene, we have employed a RMSD fitting procedure to map the dynamics to the molecular frame of reference, the Eckhart frame, instead of using internal coordinates for this procedure.

Figure 8 displays the IR intensities of the individual stretching modes of the isotopomers as generated by the extended GNC procedure. Their labels have been adopted from the nomenclature introduced in ref 24 and are listed in Table 2 together with the contributions of the leading stretching components for the reader's convenience. In $\text{CH}_n\text{D}_{5-n}^+$ the stretching modes are localized on either of the two subunits, i.e., the tripod or H_2/HD moiety. Thus, the modes are labeled either "m" for the two H_2/HD moiety stretches or "t" for the three tripod stretches. The subscript H or D indicates whether a proton or a deuteron stretch dominates the mode, whereas the superscript refers to the symmetry of the mode: "a", antisymmetric; "s", symmetric; "i", in-plane; and "u", uncoupled.

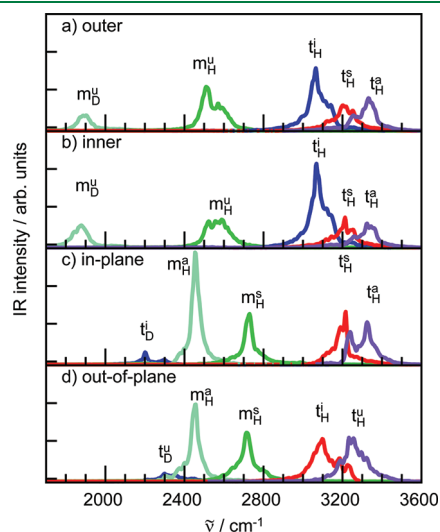


Figure 8. Spectral analysis of the stretching region for the four different isotopomers of the CH_4D^+ isotopologue (recall that the two out-of-plane isotopomers yield identical spectra). The disentangled spectral contributions are depicted in panel a for the outer isotopomer, b for inner, c for in-plane, and d for out-of-plane.

In pure CH_5^+ the tripod modes consist of the antisymmetric and symmetric out-of-plane stretches (t_H^i and t_H^s) and the in-plane stretch (t_H^i), and the H_2 moiety modes are the symmetric and antisymmetric stretches, m_H^s and m_H^a , respectively, see Supporting Information to ref 24 for a pictorial description of the corresponding atomic motions. Upon partial deuteration in the CH_4D^+ case, the symmetries of the CH_5^+ modes may be lowered, and formerly (anti)symmetric modes may now "uncouple" to yield only one dominant stretch. For instance, if the deuteron substitutes a proton in the H_2 moiety, the two former H_2 moiety modes, m_H^a and m_H^s , become m_D^u and m_H^u modes at low and high frequencies, respectively, see Figure 8a,b and Table 2. Notably, both the spectral contributions and the normal modes of the tripod stretches therein remain nearly the same and, in addition, are similar to CH_5^+ (data not shown); t_H^a has the highest frequency followed by t_H^s and t_H^i . Note also that the tripod and HD moiety stretches are not coupled, supporting the adequacy of the concept of two principle structural building blocks of protonated methane species: " H_2 moiety" and " CH_3 tripod".

When the deuteron resides in the tripod, the H_2 moiety modes are similar to those of bare CH_5^+ ; note that here the symmetric stretch m_H^s is higher in frequency than the antisymmetric stretch m_H^a , see Figure 8c,d and Table 2. However, the coefficients of t_D^i and t_D^u in Table 2 indicate that there is a noticeable coupling between the H_2 moiety stretches H1 and H2 to the tripod deuteron stretches D3 and D4, as expected from the similarity of their frequencies. When the deuteron occupies the in-plane position within the tripod, Figure 8c, the in-plane stretch, t_D^i , simply red-shifts, whereas the other two stretching modes t_H^s and t_H^a of the out-of-plane protons retain their symmetric and antisymmetric character similar to isotopomers characterized by a HD moiety or to bare CH_5^+ . Finally, the symmetry is again lowered when the deuteron enters one of the out-of-plane sites, see Figure 8d. In this case the symmetric and antisymmetric tripod modes decouple into t_D^u and t_H^u modes, and t_D^u moves to the region at about 2300 cm^{-1} while t_H^u replaces t_H^a in its frequency range. Note that the stretch coefficients of the out-of-plane modes listed in Table 2 differ by at most 0.02 between the two out-of-plane isotopomers. Furthermore, these coefficients are almost identical if they are determined on a much smaller data set of only 60 trajectories, i.e., the data basis that was used in ref 24, which reflects the good convergence of the GNCs even if the statistics are limited. This discussion makes clear the point that a trustworthy assignment of the IR spectrum of an utmost floppy molecule such as CH_4D^+ not only requires one to split its total spectrum properly into the contributions stemming from the five isotopomers but also requires a full vibrational analysis of the latter by the extended GNC approach.

Thus, our detailed analysis confirms the capability of the extended GNC technique to treat highly nontrivial, truly multiple reference structure cases like the $\text{CH}_n\text{D}_{5-n}^+$ isotopologues of protonated methane.

Table 2. Dominant Stretch Components HX and DX to the Stretching Modes of the Isotopomers As Indicated

outer		inner		in-plane		out-of-plane	
m_D^u	0.87 D1 – 0.29 H2	m_D^u	0.90 D2 – 0.28 H1	t_D^i	0.93 D3 + 0.36 H1	t_D^u	0.91 D4 + 0.37 H2
m_H^u	0.93 H2 + 0.12 D1	m_H^u	0.91 H1 + 0.18 H3	m_H^a	0.77 H2 – 0.55 H1	m_H^a	0.68 H1 – 0.64 H2
t_H^i	0.97 H3 + 0.16 H5/H4	t_H^i	0.95 H3 + 0.20 H5/H4	m_H^s	0.68 H1 + 0.54 H2	m_H^s	0.62 H1 + 0.60 H2
t_H^s	0.70 H4 + 0.68 H5	t_H^s	0.70 H5 + 0.67 H4	t_H^s	0.71 H5 + 0.69 H4	t_H^i	0.95 H3 – 0.17 H1
t_H^a	0.70 H5 – 0.68 H4	t_H^a	0.70 H4 – 0.67 H5	t_H^a	0.70 H4 – 0.68 H5	t_H^u	0.97 H5 – 0.14 H3

5. CONCLUSIONS AND OUTLOOK

We have presented a significant extension of the generalized normal coordinate (GNC) analysis method introduced recently, which now enables the algorithm to handle multiple reference structures for a comprehensive vibrational analysis of molecular dynamics trajectories. These reference structures correspond to local minima of the PES, which either define chemically different conformations of the molecule or represent chemically equivalent structures that result from a permutation of atoms of the same species.

For each and every time frame of the trajectories, the probabilities of the molecule to occupy either of these reference structures are computed. Thereby, the trajectories are split into the contributions of the references. A time correlation formalism which weights the trajectory frames with the computed probabilities yields split spectra that represent the contributions of the underlying conformations to the total spectrum. Moreover, the scheme is capable to resolve the permutational symmetries, which permits one to compute GNCs for each conformation and, thereby, to assign the peaks in the conformation-specific IR spectra to atomic motion.

We have demonstrated the methodology in detail for two selected spectroscopic problems. The first task was to cleanly assign the vibrational modes of the methyl group of isoprene, which rotates nearly freely at ambient conditions. In paper I, it was shown that these modes are unsatisfactorily resolved when one assumes a quasi-rigid molecule as in the original GNC scheme. However, the extended GNC approach introduced here, employing three reference structures representing the permutational symmetry of the methyl group, fully resolves the vibrational bands, and the GNCs of these modes comply with the C_s symmetry of the molecule.

The second, much more demanding task was the comprehensive vibrational analysis of the CH_4D^+ isotopologue of protonated methane, the latter being widely considered to be among the most prominent and a truly challenging representatives of the class of floppy molecules. Here, 120 reference structures were needed to resolve the spectral contributions of the five isotopomers (“conformations”), each covering 24 permutational symmetries. The analysis showed that, although the dynamics of switching between the isotopomers occurs on a time scale as fast as 60 fs, the algorithm is able to split the computed total IR spectrum into physically reasonable contributions of the conformational states, i.e., into five isotopomer-specific IR spectra to be assigned separately using GNCs.

Having assigned the IR spectrum of a most challenging case such as partially deuterated protonated methane in terms of atomic motions, we expect the extended GNC method to be useful for the plethora of cases where a single equilibrium or reference structure is not sufficient to understand vibrational spectra. Last but not least, the very idea of using a dynamical projection scheme to split trajectories and thus spectra—a posteriori for analysis and not a priori for generation—can be readily transferred to other analysis schemes of response properties of molecular systems which are based on time correlation functions generated by (ab initio) molecular dynamics underlying theoretical spectroscopy in the Heisenberg picture.

AUTHOR INFORMATION

Corresponding Author

*E-mail: gerald.mathias@physik.uni-muenchen.de.

Present Addresses

[†]Lehrstuhl für BioMolekulare Optik, Ludwig-Maximilians-Universität München, Oettingenstr. 67, 80538 München, Germany

[§]Quantum Dynamics Group, Institut für Physik, Universität Rostock, 18051 Rostock, Germany

^{||}PULSE Institute and Department of Chemistry, Stanford University, Stanford, California 94305, United States

[⊥]Chemical and Materials Science Division, Pacific Northwest National Laboratory, P.O. Box 999, Richland, Washington 99352, United States

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

We are grateful to T. Zelleke for his contributions to the implementation, to DFG (MA 1547/4 to D.M. and SFB 749/C4 to G.M.) and FCI (Chemiefonds—Stipendium to A.W.) for financial support, as well as to HLRB II (München), BOVILAB@RUB (Bochum), and RV-NRW (Dortmund) for computational resources.

REFERENCES

- (1) Siebert, F. *Methods Enzymol.* **1995**, *246*, 501–526.
- (2) Vogel, R.; Siebert, F. *Curr. Opin. Chem. Biol.* **2000**, *4*, 518–523.
- (3) Bieske, E. J.; Dopfer, O. *Chem. Rev.* **2000**, *100*, 3963–3998.
- (4) Barth, A.; Zscherp, C. *Q. Rev. Biophys.* **2002**, *35*, 369–430.
- (5) Barth, A. *Biochim. Biophys. Acta* **2007**, *1767*, 1073–1101.
- (6) Polfer, N. C.; Oomens, J. *Mass Spectrom. Rev.* **2009**, *28*, 468–494.
- (7) Marx, D.; Hutter, J. *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods*; Cambridge University Press: Cambridge, U. K., 2009; pp 11–75.
- (8) Silvestrelli, P. L.; Bernasconi, M.; Parrinello, M. *Chem. Phys. Lett.* **1997**, *277*, 478–482.
- (9) Nonella, M.; Mathias, G.; Eichinger, M.; Tavan, P. *J. Phys. Chem. B* **2003**, *107*, 316–322.
- (10) Vogel, R.; Siebert, F.; Mathias, G.; Tavan, P.; Fan, G.; Sheves, M. *Biochemistry* **2003**, *42*, 9863–9874.
- (11) Gaigeot, M.-P.; Sprik, M. *J. Phys. Chem. B* **2003**, *107*, 10344–10358.
- (12) Rousseau, R.; Kleinschmidt, V.; Schmitt, U. W.; Marx, D. *Angew. Chem., Int. Ed.* **2004**, *43*, 4804–4807.
- (13) Gaigeot, M.-P.; Vuilleumier, R.; Sprik, M.; Borgis, D. *J. Chem. Theory Comput.* **2005**, *1*, 772–789.
- (14) Iftimie, R.; Tuckerman, M. E. *J. Chem. Phys.* **2005**, *122*, 214508.
- (15) Asvany, O.; Kumar, P. P.; Redlich, B.; Hegemann, I.; Schlemmer, S.; Marx, D. *Science* **2005**, *309*, 1219–1222.
- (16) Martinez, M.; Gaigeot, M.-P.; Borgis, D.; Vuilleumier, R. *J. Chem. Phys.* **2006**, *125*, 144106/14.
- (17) Kumar, P.; Marx, D. *Phys. Chem. Chem. Phys.* **2006**, *8*, 573–586.
- (18) Gaigeot, M.-P.; Martinez, M.; Vuilleumier, R. *Mol. Phys.* **2007**, *105*, 2857–2878.
- (19) Mathias, G.; Marx, D. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 6980–6985.
- (20) Masia, M.; Forbert, H.; Marx, D. *J. Phys. Chem. A* **2007**, *111*, 12181–12191.
- (21) Baer, M.; Mathias, G.; Kuo, I.-F. W.; Tobias, D. J.; Mundy, C. J.; Marx, D. *ChemPhysChem* **2008**, *9*, 2703–2707.
- (22) Cimas, A.; Vaden, T. D.; de Boer, T. S. J. A.; Snoek, L. C.; Gaigeot, M. P. *J. Chem. Theory Comput.* **2009**, *5*, 1068–1078.
- (23) Thomas, V.; Iftimie, R. *J. Phys. Chem. B* **2009**, *113*, 4152–4160.
- (24) Ivanov, S. D.; Asvany, O.; Witt, A.; Hugo, E.; Mathias, G.; Redlich, B.; Marx, D.; Schlemmer, S. *Nat. Chem.* **2010**, *2*, 298–302.
- (25) Heyden, M.; Sun, J.; Funkner, S.; Mathias, G.; Forbert, H.; Havenith, M.; Marx, D. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 12068–12073.

- (26) Zhang, C.; Donadio, D.; Galli, G. *J. Phys. Chem. Lett.* **2010**, *1*, 1398–1402.
- (27) Baer, M.; Marx, D.; Mathias, G. *Angew. Chem., Int. Ed.* **2010**, *49*, 7346–7349.
- (28) Witt, A.; Ivanov, S. D.; Mathias, G.; Marx, D. *J. Phys. Chem. Lett.* **2011**, *2*, 1377–1381.
- (29) Huang, X. C.; Carter, S.; Bowman, J. *J. Chem. Phys.* **2003**, *118*, 5431–5441.
- (30) Huang, X. C.; Braams, B. J.; Bowman, J. M. *J. Chem. Phys.* **2005**, *122*, 044308.
- (31) Bowman, J. M.; Carrington, T.; Meyer, H.-D. *Mol. Phys.* **2008**, *106*, 2145–2182.
- (32) Jin, Z.; Braams, B.; Bowman, J. *J. Phys. Chem. A* **2006**, *110*, 1569–1574.
- (33) Park, M.; Shin, I.; Singh, N. J.; Kim, K. S. *J. Phys. Chem. A* **2007**, *111*, 10692–10702.
- (34) Kaledin, M.; Kaledin, A. L.; Bowman, J. M.; Ding, J.; Jordan, K. D. *J. Phys. Chem. A* **2009**, *113*, 7671–7677.
- (35) Baer, M.; Marx, D.; Mathias, G. *ChemPhysChem* **2011**, *12*, 1906–1915.
- (36) Gordon, R. G. *Adv. Magn. Reson.* **1968**, *3*, 1–42.
- (37) Berens, P. H.; Wilson, K. R. *J. Chem. Phys.* **1981**, *74*, 4872–4882.
- (38) Putrino, A.; Parrinello, M. *Phys. Rev. Lett.* **2002**, *88*, 176401.
- (39) Wheeler, R. A.; Dong, H.; Boesch, S. E. *ChemPhysChem* **2003**, *3*, 382–384.
- (40) Wheeler, R. A.; Dong, H. *ChemPhysChem* **2003**, *4*, 1227–1230.
- (41) Schmitz, M.; Tavan, P. *J. Chem. Phys.* **2004**, *121*, 12233–12246.
- (42) Schmitz, M.; Tavan, P. *J. Chem. Phys.* **2004**, *121*, 12247–12258.
- (43) Agostini, F.; Vuilleumier, R.; Ciccotti, G. *J. Chem. Phys.* **2011**, *134*, 084302.
- (44) Mathias, G.; Baer, M. *J. Chem. Theory Comput.* **2011**, *7*, 2028–2039.
- (45) Towns, T. G.; Carreira, L. A.; Irwind, R. M. *J. Raman Spectrosc.* **1981**, *11*, 487–492.
- (46) Marx, D.; Parrinello, M. *Nature* **1995**, *375*, 216–218.
- (47) Marx, D.; Parrinello, M. *Science* **1999**, *284*, 59–61.
- (48) McCoy, A. B.; Braams, B.; Brown, A.; Huang, X.; Jin, Z.; Bowman, J. *J. Phys. Chem. A* **2004**, *108*, 4991–4994.
- (49) Huang, X.; Johnson, L.; Bowman, J.; McCoy, A. *J. Am. Chem. Soc.* **2006**, *128*, 3478–3479.
- (50) Huang, X.; McCoy, A. B.; Bowman, J. M.; Johnson, L. M.; Savage, C.; Dong, F.; Nesbitt, D. J. *Science* **2006**, *311*, 60–63.
- (51) Johnson, L.; McCoy, A. *J. Phys. Chem. A* **2006**, *110*, 8213–8220.
- (52) Wanga, X.-G.; Carrington, T., Jr. *J. Chem. Phys.* **2008**, *129*, 234102.
- (53) Hinkle, C. E.; Petit, A. S.; McCoy, A. B. *J. Mol. Spectrosc.* **2011**, *268*, 189–198.
- (54) “Isotopologue” (doi: 10.1351/goldbook.I03351) and .isotopomer. (doi: 10.1351/goldbook.I03352) are used according to IUPAC, see <http://goldbook.iupac.org> (created by Nic, M.; Jirat, J.; Kosata, B.; updates compiled by Jenkins, A.; accessed 11/2011).
- (55) Muller, H.; Kutzelnigg, W.; Noga, J.; Klopper, W. *J. Chem. Phys.* **1997**, *106*, 1863–1869.
- (56) Ramírez, R.; Lopez-Ciudad, T.; Kumar, P.; Marx, D. *J. Chem. Phys.* **2004**, *121*, 3973–3983.
- (57) Wilson, E.; Decius, J. C.; Cross, P. C. *Molecular Vibrations*; McGraw Hill: New York, 1955; pp 54–76.
- (58) Cremer, D.; Pople, J. *J. Am. Chem. Soc.* **1975**, *97*, 1354–1358.
- (59) VandeVondele, J.; Krack, M.; Mohamed, F.; Parrinello, M.; Chassaing, T.; Hutter, J. *Comput. Phys. Commun.* **2005**, *167*, 103–128.
- (60) Hutter, J.; Alavi, A.; Deutsch, T.; Bernasconi, M.; Goedecker, S.; Marx, D.; Tuckerman, M.; Parrinello, M. CPMD: Car–Parinello Molecular Dynamics, version 3.10; IBM Corp: Endicott, NY, 1990; MPI für Festkörperforschung Stuttgart: Stuttgart, Germany, 1997. www.cpmid.org (accessed Nov. 2011).
- (61) Hsu, S.; Kemp, M.; Pochan, J.; Benson, R.; Flygare, W. *J. Chem. Phys.* **1969**, *50*, 1482–1483.

The Absorption Spectrum of Cytosine Tautomers: Beyond the Static Approach

Alex Domingo,^{*,†} Antonio Rodríguez-Forteza,[†] and Coen de Graaf^{*,†,‡}

[†]Departament de Química Física i Inorgànica, Universitat Rovira i Virgili, Marcel·lí Domingo s/n, 43007 Tarragona, Spain

[‡]Institució Catalana de Recerca i Estudis Avançats, Pg. Lluís Companys 23, 08010 Barcelona, Spain

ABSTRACT: The absorption spectrum of cytosine in water has been studied by combining Car–Parrinello molecular dynamics (MD) with a multiconfigurational perturbation theory treatment of the electronic structure. The MD simulations were performed for four different tautomeric forms of cytosine in a unit cell with 60 water molecules. The relative energies and transition dipole moments of a large number of excited states have been calculated on a representative sample of conformations along the MD trajectories. In this way, the broad experimental peaks can be decomposed, and the effect of the distortions on the nature of the excited states can be assessed. The loss of planarity of the molecule is significant, and hence, the excited states can no longer be defined as pure $n \rightarrow \pi^*$ or $\pi \rightarrow \pi^*$ excitations. We propose an analysis to assign the different transitions according to the main contribution. The keto N1H form turns out to be the most stable one, and the calculated spectra of this tautomer show good agreement with experimental measurements. The mixed $n\pi^*/\pi\pi^*$ character of some states leads to a significant increase of intensity in spectral regions dominated by the dark $n\pi^*$ transitions considering a planar structure.

1. INTRODUCTION

The sensibility of the DNA molecule to UV light absorption can lead to mutations of the genetic material and eventually carcinogenesis. However, the DNA molecule has effective mechanisms of deactivation to avoid genetic damage. The constituent nucleobases of DNA have a prominent role in this process because not only do all four bases (adenine, thymine, guanine, and cytosine) have large absorption coefficients in the UV range but also the monomers and base pairs show intrinsic molecular fast decays upon electronic excitation.^{1,2}

Even though the four nucleobases are relatively simple molecules, the understanding of their electronic structure is complex. The DNA base cytosine is known to have multiple decay pathways that involve various excited states. The UV radiation of cytosine produces primarily dipole allowed transitions to singlet $\pi\pi^*$ states, the so-called bright states, and dipole forbidden transitions to singlet $n\pi^*$ states at a lower rate, the so-called dark states. Theoretical and experimental research on the potential energy surfaces (PES) of the ground state and the lowest excited states of cytosine have found many plausible radiationless deactivation mechanisms for this chromophore. Nowadays, the proposed decay pathways for cytosine include multiple internal conversion via conical intersections (CI) at different regions of the hyper-PES between various $^1\pi\pi^*$ and $^1n\pi^*$ states and the ground state.^{3–7} A relaxation mechanism through an intersystem crossing (ISC) to the lowest $^3\pi\pi^*$ state has been also suggested.⁸ Furthermore, some decay paths involving conformational distortions⁹ or excited tautomerization¹⁰ pointed out the significant role of the structural flexibility of cytosine. These results prove the large versatility and high efficiency of the cytosine molecule to decay back to the ground state but, additionally, the high complexity of its electronic structure.

At first sight, the absorption spectra of cytosine are relatively simple. They show basically two intense and broad absorption bands with some additional minor features depending on the

experimental conditions.^{11–14} A brief summary of available spectroscopic data for cytosine is shown in Table 1. The main bands of the absorption spectra of cytosine correspond to transitions I (~ 4.5 eV) and IV (~ 6.0 eV), which are transitions in the molecular plane assigned to $\pi\pi^*$ excitations.¹⁵ Transitions II and III (5–6 eV) are usually masked by the more intense $\pi\pi^*$ bands, with their characterization by means of experimental polarization data and transition moments becoming difficult.^{16–18}

The theoretical research of the electronic structure has put some light into these weak absorptions. Early calculations of the cytosine molecule using an all valence electrons self consistent field molecular orbital with configuration interaction (SCF-MO–CI) resulted in detailed information about the characteristics of both the $\pi\pi^*$ ²⁰ and $n\pi^*$ ²¹ transitions, identifying transition II at 5.2 eV as an $n\pi^*$ excitation from the nonbonding orbital of the carbonyl group. Moreover, Matos and Roos used the complete active space self consistent field (CASSCF) method to treat the complete π electronic system of cytosine. They found that the lone pairs of the O and N atoms have a large effect on the electron correlation of the molecule and calculated the three lowest $\pi\pi^*$ transition energies to be 5.6, 6.9, and 8.1 eV.¹⁷ Subsequent calculations by Fülischer and Roos included σ – π polarization through the second-order perturbation theory using a CASSCF reference wave function (CASPT2) to treat the dynamical correlation of σ electrons. They obtained more accurate results with this method and calculated the transition energies of four $\pi\pi^*$ states at 4.4, 5.4, 6.2, and 6.7 eV, a $n\pi^*$ transition as the second excited state at 5.0 eV, and a possible higher $n\pi^*$ transition at 6.5 eV.²² Petke et al. performed multi-reference CI calculations of cytosine that led to similar conclusions for the lowest transitions but pointed at the complexity of the higher energy region due to spectral congestion above 5.2 eV.¹⁸

Received: October 3, 2011

Published: December 05, 2011

Table 1. Summary of Spectroscopic Data for Cytosine

spectrum medium	ref	transition energy (eV)					
		I	II	III	IV	V	VI
trimethyl phosphate vapor	11	4.5	5.2		6.1, 6.7		
water (pH = 7)	12	4.3					
water (pH = 2)	13	4.7					
crystal monohydrate	14	4.5			5.9		
water ^d	16	4.7	5.3	5.6	6.2	7.7	8.1
ethanol	19	4.5	5.2		6.0		

^d Transitions I–IV adjusted from crystal polarized reflection spectra. Transitions V–VI estimated from thin film spectrum.

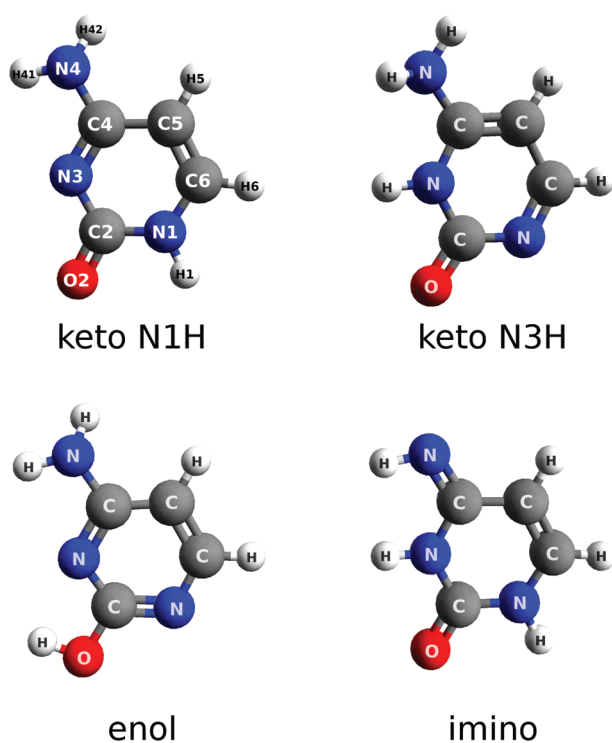


Figure 1. Structures of four tautomeric forms of cytosine.

The influence of different solvents to the electronic structure of cytosine has been another important subject of research. The absorption spectrum of cytosine depends on the dielectric constant of the solvent (Table 1) and, in the case of water, on the pH value.^{13,23} Furthermore, cytosine has various tautomeric forms that can be classified on the basis of their functionalization in three basic structures: keto, enol, and imino (Figure 1). The population of these tautomers varies depending on the characteristics of the solvent, and additionally, they have quite different electronic properties, complicating even more the interpretation of spectroscopic data.²⁴ Even though various tautomeric structures can be present for a given solvating condition, the enol form is the most stable in apolar solvents and the gas phase, while the keto form is the most stable in polar solvents.^{19,25–29}

There has been much effort to incorporate aqueous solvent effects into the theoretical models of cytosine and other organic molecules to achieve biologically relevant data.³⁰ To simulate the

electrostatic effect of the solvent in the computational model, one commonly includes some kind of continuum reaction-field around the quantum system.^{31–35} Additionally, some explicit water molecules can be included in the electronic structure calculation to achieve a more accurate description of the solute–solvent interactions.^{36–39} The main known effect of aqueous solvent on the electronic structure of cytosine is a blue shift of the $n-\pi^*$ transitions due to stabilization of the electron lone pairs, which pushes the lower dark excitations to the spectral congestion region.

The cytosine tautomer found in the Watson and Crick structure of DNA is the keto form.⁴⁰ However, the large capability of the nucleobases to transfer protons and the charge transfer phenomena through the nucleobases stacks in the DNA chain^{41,42} can lead to tautomerization of the DNA bases. It has been suggested that the tautomerization of cytosine occurs primarily through ionic structures²⁵ and that photoinduced tautomerization could offer extra pathways for deactivation.^{10,36} Moreover, work on other biologic systems like urocanic acid showed the role that different tautomers can have on the absorption spectrum.⁴³ Thus, it becomes important to consider not only the keto form of cytosine but also its other tautomeric structures to understand the electronic properties of cytosine related biologic systems.

The main drawback of the theoretical models described so far is that they are limited to a static description of the system. Although it is possible to treat different tautomers and include solvent effects in the calculation, only ideal structures of the molecule have been studied so far. However, cytosine is known to be very flexible and easily loses its planarity. Molecular dynamics simulations using the Car–Parrinello scheme explored the conformational space of the DNA bases showing a wide range of accessible distortions for this molecule.⁴⁴ Thus, once a cytosine molecule deviates from planarity, it is not strictly possible to talk about π and n states and $\pi\pi^*$ and $n\pi^*$ transitions, because σ , π , and n MOs start to mix, and the interpretation of the single contributions to the absorption spectra becomes difficult to perform. Nevertheless, the inclusion of the vibro-rotational degrees of freedom on the theoretical study of these compounds is required to achieve a good understanding of their electronic structure. Nowadays, it is becoming a rule to combine molecular dynamics simulations with QM/MM calculations to study biologic systems,^{45–49} and the cytosine molecule is no exception.^{6,7,50,51} Hence, this theoretical framework incorporates the conformational evolution of the system at a given temperature plus the effect of the solvent and sophisticated electronic structure calculations.

The aim of the present work is to continue developing the *ab initio* theoretical description of the cytosine molecule by combining Car–Parrinello molecular dynamics (Car–Parrinello MD)⁵² with the complete active space self-consistent field second-order perturbation theory (CASPT2).^{53,54} We study the conformational space of four tautomeric forms of cytosine in water (Figure 1) by means of Car–Parrinello MD simulations. Subsequently, on a representative sample of the resulting trajectories, we perform CASPT2 calculations to obtain an accurate description of the electronic structure of each conformation. The solvent effects are incorporated on the CASPT2 step through two solvent models, the polarizable continuum model (PCM)⁵⁵ and a modification of PCM including some explicit water molecules. The resulting data set for each tautomer is combined to generate its absorption spectrum in water. We offer a detailed

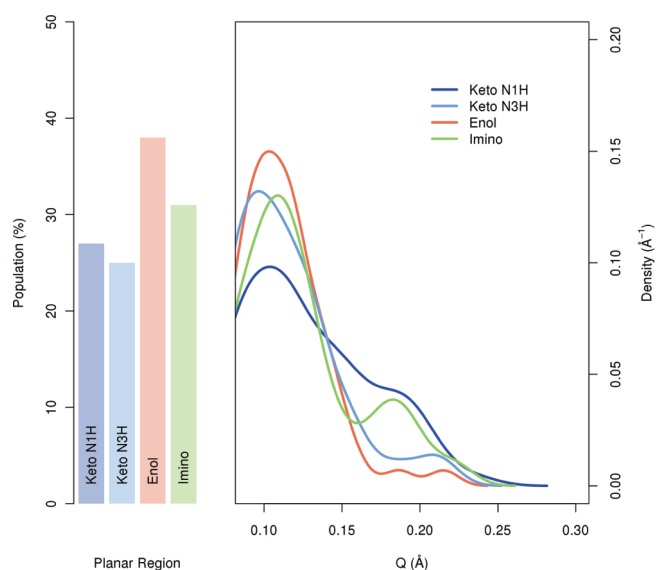


Figure 2. Degree of puckering of the cytosine aromatic ring for each tautomeric form. Left: Proportion of planar structures. Right: Value of the Q parameter for nonplanar structures.

view of the underlying structure of the absorption spectrum of cytosine, comparing the two solvent models and analyzing the large mixing observed between the different transition types.

2. COMPUTATIONAL SCHEME

In contrast to classical molecular dynamics simulations, first-principles (also called *ab initio*) molecular dynamics methods compute the forces acting on the nuclei from electronic structure calculations (typically using the density functional theory) that are performed as the molecular dynamics trajectory is generated (“on the fly”). Consequently, those systems with electronic structures that change significantly (i.e., formation/breaking of bonds) during the dynamics can be easily handled by molecular dynamics. The method developed by Car and Parrinello⁵² uses fictitious dynamics to develop the electronic orbital functions, which are only minimized at the beginning of the dynamics, preventing the need for a costly self-consistent iterative minimization at each time step. Details about the Car–Parrinello MD method and first-principles molecular dynamics in general can be found in ref 56 and references therein.

We performed Car–Parrinello MD simulations for each tautomeric form of cytosine considered in this work, namely the keto N1H, keto N3H, enol, and imino forms (Figure 1). All simulations were performed in a box with side lengths of 12.5487 Å containing 60 water molecules and one cytosine molecule. The molecular dynamics conditions were set according to the previous work by Isayev et al.⁴⁴ on cytosine in a vacuum. The core electrons are described by Troullier–Martins normconserving pseudopotentials.⁵⁷ The electronic potential was calculated by means of the BLYP density functional^{58,59} using a plane waves basis set with a cutoff of 80 Ry. The system was first equilibrated at 300 K and then maintained at a steady temperature with a Nosé–Hoover chain thermostat.^{60,61} All H atoms are deuterium isotopes, to avoid the high frequency hydrogen stretchings and be able to use a larger time step of 0.121 fs with a fictitious electron mass $\mu = 700$ au to improve the computational cost.⁶² For each tautomeric form, 100 conformational structures were

Table 2. Classification in the Six Symmetrical Distorted and Planar Forms of the 100 Conformational Structures of the Four Cytosine Tautomers

symmetrical form	relative proportion (%)			
	keto N1H	keto N3H	enol	imino
planar	27	25	38	31
chair	2	3	1	5
half-chair	13	5	7	12
envelope	14	17	13	14
screw-boat	18	19	12	12
boat	13	14	20	16
twist-boat	13	17	9	10

extracted from a trajectory of 2 ps. The MD simulations were carried out with the CPMD software version 3.11.1.⁶³

The 400 conformers selected from the Car–Parrinello MD simulations are the ones used in the electronic structure calculations of the excited states of cytosine. To ensure the sample representativeness of this set of structures, we analyzed the degree of puckering of their aromatic ring in terms of the Cremer–Pople (CP) parameters,⁶⁴ which offer a systematic method to measure the distortion of each conformer and classify them following Boeyen’s scheme.⁶⁵ All structures are classified in symmetrical nonplanar conformations based on their CP parameters Q , θ , and ϕ . The Q parameter serves as a measure of ring planarity, the more planar structures being the ones with a smaller Q . The conformers with a weighted average torsion angle smaller than $\pm 5^\circ$ are considered planar, which corresponds to Q values smaller than 0.1 Å. Figure 2 shows the relative population of planar structures and the pronounced loss of planarity of the four tautomeric forms of cytosine. Only three conformations out of 10 can be considered planar. The degree of distortion of the four tautomers is very similar, and all of them have significantly smaller Q values in water than in the gas phase.⁴⁴ Furthermore, the tautomeric forms with a protonated N3 ring atom show slightly larger average distortions. This behavior could be ascribed to a larger steric repulsion caused by the N3 hydrogen atom on the out-of-ring heteroatoms. The other two parameters θ and ϕ are used to identify the type of distortion of the conformer and assign them to one of the six symmetrical forms of a molecule with a six-membered ring. Table 2 analyzes the proportion of the symmetrical forms between the different tautomers of cytosine. The percentage of chair like forms is low for all tautomers compared with the planar and boat like forms, which are the most abundant. The presented analysis was performed using the PLATON package.⁶⁶

The aqueous solvent effects in the electronic structure calculations were computed by means of a PCM to reproduce the electrostatic interaction of water with the cytosine molecule. The electronic response of the solvent upon absorption is partially defined by the ground state properties of the solute, the so-called slow component, and the excited final state, the so-called fast component. Moreover, we performed a second set of calculations for all of the structures including in the wave function some explicit water molecules surrounding the cytosine molecule, in addition to the PCM wrapping. Hence, the effect of hydrogen bonding between the solute and closer water molecules can be incorporated into the electronic structure. Those explicit water molecules come directly from the Car–Parrinello MD, maintaining their relative orientation

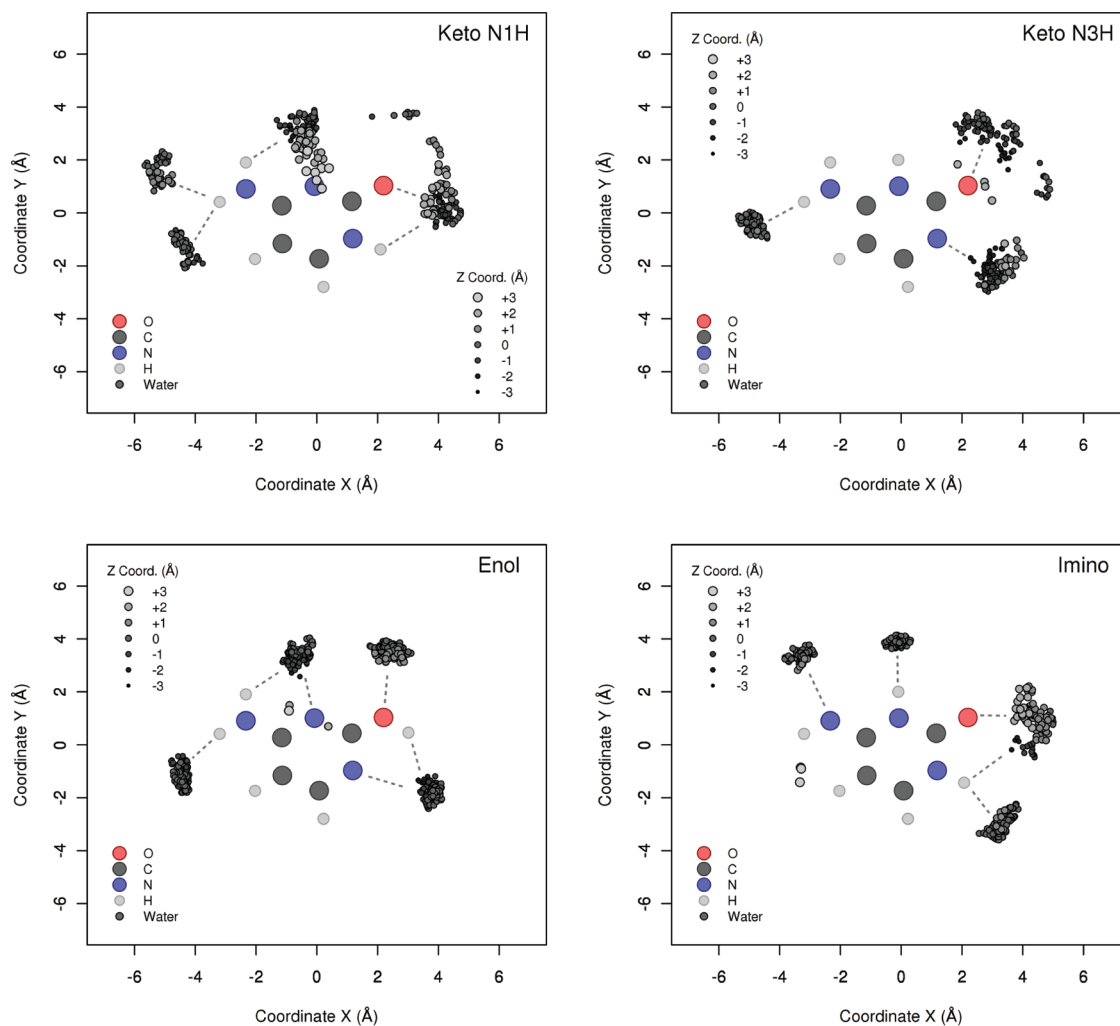


Figure 3. Relative orientation of the selected water molecules on each tautomeric form of cytosine: keto N1H (top left), keto N3H (top right), enol (bottom left), and imino (bottom right). Each graph is a superposition of the position on the molecular plane of all water molecules chosen along the 100 conformations (dark dots). The H-bonding formed by the solvent molecules is qualitatively represented by dashed lines to the respective ideal tautomeric structure.

to the cytosine conformer and being selected on the basis of their distance to the electronegative atoms of the molecule.

The two keto tautomers have three water molecules added, whereas the enol and imino tautomers have four explicit water molecules. This choice is based on the number of available sites for H-bonding. We consider the interaction of water with the functional groups and the heteroatoms of the ring. Therefore, the keto tautomers have three main interaction sites with the solvent, the carbonyl and amine groups and one lone pair of a N atom on the ring. In the case of the enol and imino tautomers, we add an additional water to maintain the balance between the two N atoms of the ring. Figure 3 shows the relative orientation of the selected explicit water molecules on the cytosine tautomers. Even though each cytosine conformer has its closest water molecules particularly oriented, they are found primarily positioned near the lone pairs and the various functional groups, as expected. The superposition of the water molecules extracted for all conformers shows how the polarity of the tautomer affects the mobility of the first solvating shell. The water molecules are significantly more fixed along the MD simulation of the enol tautomer than for the less polar keto forms.

The scope of this work is restricted to the lowest 9 eV region of the absorption spectrum of cytosine, not only to cover the largely studied first and second intense broad bands but also to explore the higher energy features observed above 7 eV.¹⁶ Thus, we performed CASPT2 calculations of the 12 lowest singlet excited states of each conformer of cytosine with the two solvent models proposed, totalling 800 electronic structure calculations. The active space (CAS) of these calculations is formed by 14 electrons in the eight π orbitals plus two nonbonding σ orbitals. This CAS allows the description of the ${}^1\pi\pi^*$ and ${}^1n\pi^*$ transitions that contribute most to the absorption spectra of cytosine with an accurate treatment of the correlation from the lone pair electrons¹⁷ and the $\sigma-\pi$ polarization.²² The basis set for the cytosine molecule is of the atomic natural orbitals (ANO) type including scalar relativistic effects⁶⁷ to obtain an optimum treatment of correlation and polarization. The C, N, and O atoms of cytosine have (4s,3p,2d,1f) contracted basis functions and the H atom a (2s,1p) contracted set. The contracted basis set employed for the explicit water molecules is (3s,2p,1d) for the O atom and (2s,1p) for the H atom. These basis sets are large enough to describe all considered valence states of cytosine. We performed the

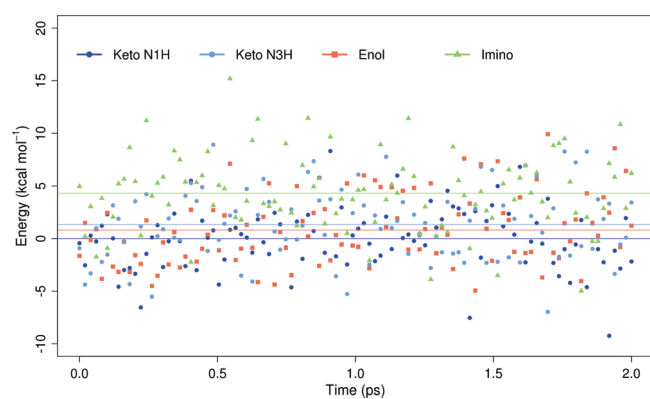


Figure 4. Ground state energy of the 400 conformers extracted from the Car–Parrinello MD simulations. Results from CASPT2 calculations using a PCM solvent model for water. The reference energy has been set to the average energy of the keto N1H tautomer. Horizontal lines represent the average energy of each tautomeric form.

CASPT2 calculations with a shift factor of 0.20 au,⁶⁸ which we previously tested to be the minimum value to eliminate possible intruder states. The intensity of the 11 lowest electronic transitions was computed by state interaction calculations between the ground state and all of the excited states at the CASPT2 level, inserting the transition dipole operator into the Hamiltonian.⁶⁹ The electronic structure calculations were carried out with the MOLCAS 7 package.^{70,71} The resulting data from the 800 calculations were statistically treated with the R package⁷² to generate the corresponding spectra by means of kernel density estimations with Gaussian kernel functions of 0.13 ± 0.01 eV of bandwidth (bw), depending on the characteristics of the data set.

Computer codes have been developed to automate the extraction of the target molecule(s) from the MD trajectories and to transform this information into a readable format for the electronic structure code. Moreover, some scripts were made to process the large amount of data generated in the computational procedure.

3. RESULTS

We obtained the absorption spectrum of the keto N1H, keto N3H, enol, and imino tautomers in water. For each tautomer, we show two spectra, one using a PCM solvent model and another with a PCM including some explicit water molecules. Both spectra are compared and analyzed in the next sections. Additionally, to check the molecular description obtained by the combination of molecular dynamics with CASPT2 calculations, we show the energetic evolution of each tautomer along the Car–Parrinello MD simulation. The energy values of spectroscopic properties are given in electronvolts, and the analysis of the relative energy between the four tautomers is done in kilocalories per mole to ease the comparison with previous studies.

3.1. Relative Stability of Cytosine Tautomers. Figure 4 shows the CASPT2 energies of the 400 conformers selected from the Car–Parrinello MD simulations of the considered cytosine tautomers. We compared the relative energy of the electronic ground state of all of the structures using a PCM solvent model of water. We did not use the solvent model including explicit water molecules, in order to make easier the analysis. A rigorous study of the relative stability of the cytosine tautomers in water is out of the scope of the present work, because it is much more complex than what we can obtain from

the present calculations. Nonetheless, this straightforward comparison served to check that the description of the different tautomers offered by the combination of Car–Parrinello MD with CASPT2 is correct.

The tautomer with the lowest energy is the keto N1H, followed by the enol at $0.8 \text{ kcal mol}^{-1}$ higher energy, the keto N3H at $1.3 \text{ kcal mol}^{-1}$, and the imino form being the most energetic at $4.3 \text{ kcal mol}^{-1}$. These results are in good qualitative agreement with previous theoretical studies about the relative stability of cytosine tautomers in water.^{19,32,33,36} The most stable form in water is known to be the keto N1H, followed by the enol tautomer. The keto N3H and imino forms are both more unstable, but their difference in energy varies depending on the theoretical model employed. Our results show that the energy between the four tautomers ranges from 1 to 5 kcal mol^{-1} . These values are in good agreement with other theoretical studies that obtained differences in energy smaller than 10 kcal mol^{-1} using diverse methods. Such small energetic differences suggest that the barriers of the tautomerization processes of cytosine can be easily overcome at room temperature.⁷³ Actually, the energetic fluctuations of these tautomers during the MD simulation is larger than 5 kcal/mol , each of the four tautomers being the lowest-energy form at some point of the simulation.

3.2. Absorption Spectra of Cytosine Tautomers. The spectra of cytosine tautomers are shown in Figures 5–9. The global absorption spectrum has been decomposed into its individual contributions for each transition type, the $\pi\pi^*$, $n_{\text{O}}\pi^*$, and/or $n_{\text{N}}\pi^*$. Since the wave function describes the complete electronic system involved in this kind of excitation, we know in detail the characteristics of those excited states. However, the classification of transitions between the $\pi\pi^*$ and the $n\pi^*$ is not straightforward. Once the molecule becomes distorted, it is not strictly possible to talk about the symmetric features of the planar structure. The MOs can mix formal aromatic π orbitals with σ orbitals to some degree. Moreover, the MO approach employed in these calculations is intrinsically delocalized, which can result in nonbonding σ orbitals spanning over two atoms with lone electron pairs. Therefore, the breakdown of the total absorption spectrum has been done by thoroughly analyzing the contributions to the orbitals in the CAS and by a population analysis of the atoms in the excited states.

We started by identifying all of the active orbitals of each conformer by means of their atomic orbital coefficients. We focused on finding the nonbonding orbitals of the corresponding O and N atoms with lone pairs. No attempts were made to classify the remaining π orbitals. The deviations from the strict π symmetry of the planar geometry are significant, making it too complicated to discriminate between the $\pi\pi^*$ transitions. The next step consisted of distinguishing those electronic transitions that significantly change the population of the nonbonding orbitals. We set a minimum criterion for the nonbonding orbitals of 0.4 electrons for the difference between the ground state and the excited state to accept those transitions as $n\pi^*$ excitations. The rest of the transitions were assigned to $\pi\pi^*$.

It must be noted that in the nonideal conformers taken from the simulation, the composition of the 12 lowest states is not equal for each structure of a single tautomer. The proportion of $n\pi^*$ and $\pi\pi^*$ transitions varies from one structure to another, resulting in a total number of different transitions found along the series of conformers to be larger than 11. Depending on the structural distortions of the conformer, it can show either a complete lack of $n\pi^*$ transitions or more $n\pi^*$ than $\pi\pi^*$ excitations.

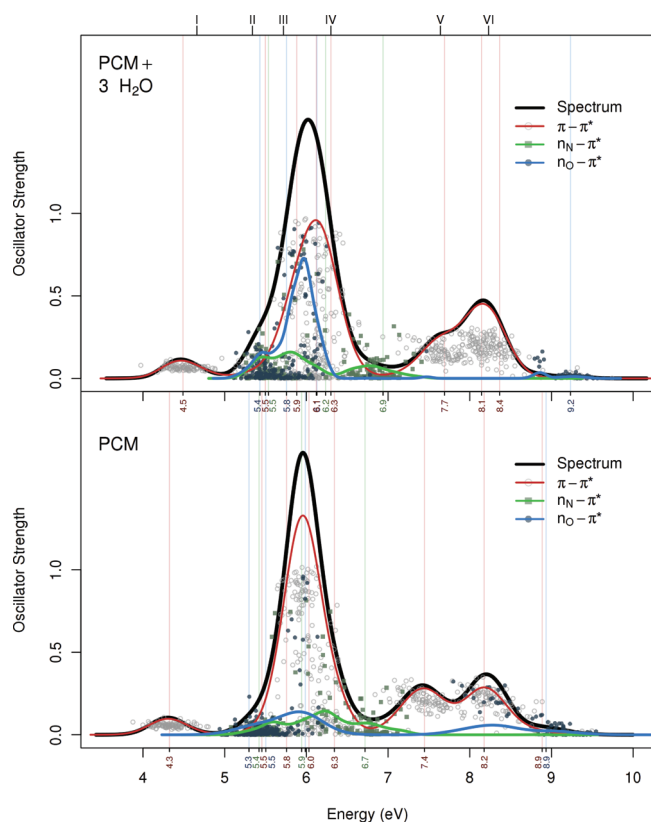


Figure 5. Computed absorption spectra of cytosine keto N1H tautomer. Result from 100 conformers using a PCM solvent model (bottom) and, additionally, with three explicit water molecules (top). Density estimation with Gaussian kernel functions ($bw = 0.14$ eV). Roman numbers on top mark the experimentally observed features.¹⁶

Keto N1H Tautomer. Figure 5 shows the computed absorption spectra of the cytosine keto N1H tautomer in water. We have identified eight $\pi\pi^*$, four $n_{\text{O}}\pi^*$, and three $n_{\text{N}}\pi^*$ transitions along all conformers of this tautomer. The number of these transitions is independent of the solvent model employed. The explicit water molecules cause a slight blue shift of ~ 0.2 eV on all transitions except for a few $\pi\pi^*$ of the high absorption band that remain stable. The major effect of the inclusion of the water molecules in the calculation is a large increase on intensity of some $n_{\text{O}}\pi^*$ excitations. These $n_{\text{O}}\pi^*$'s are located in the region of spectral congestion around ~ 6 eV, where the most intense $\pi\pi^*$ transitions occur. This characteristic suggests that the water molecules could favor the π character of these transitions, increasing the overlap between the involved states and, consequently, increasing their absorption. Additionally, the three water molecules improve the description of the highest excited states, reducing the mixing between the $\pi\pi^*$ and $n_{\text{O}}\pi^*$ transitions observed at 9 eV in the PCM spectrum.

There are four pure $\pi\pi^*$ bands in the spectrum of the keto N1H, the first absorption band located at 4.5 eV and the two peaks at 7.7 and 8.1 eV, which mask a fourth transition at 8.4 eV. The other spectral features that can be unequivocally assigned to single excitations are two optically nearly inactive transitions, a $n_{\text{N}}\pi^*$ at 6.9 eV and a $n_{\text{O}}\pi^*$ at 9.2 eV. The remaining electronic transitions appear in the high absorption band at ~ 6 eV, which is formed by a mix of various absorptions. The transitions observed can be grouped into three main bands that correlate rather well

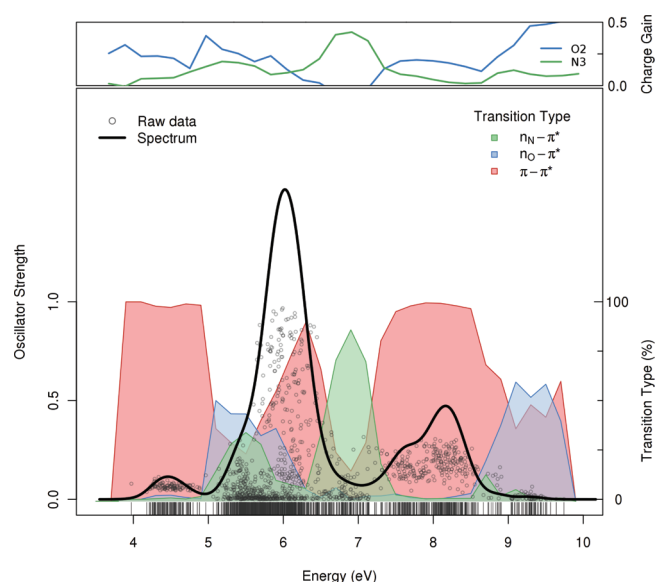


Figure 6. Analysis of the absorption spectra of cytosine keto N1H tautomer with three explicit water molecules. The colored graphs behind the spectrum show the percentage in intervals of 0.2 eV of each type of transition, and the graph on the upper part shows the LoProp population analysis on the O and N atoms with lone pairs.

with the bands identified by Zaloudek et al.¹⁶ Transition II has contributions of all three types of excitations. Transition III is quite intense and formed equally by $\pi\pi^*$ and $n_{\text{O}}\pi^*$ absorptions, and transition IV is raised basically by two $\pi\pi^*$, which seem to correlate to the E_{1u} state of benzene, as suggested by Tinoco and Clark.¹¹ Furthermore, transition I clearly corresponds to the first $\pi\pi^*$ band at 4.5 eV. In the higher energy region, we observe three $\pi\pi^*$ transitions that match very accurately in energy and band morphology to transitions V, formed by one $\pi\pi^*$, and transition VI, formed by two $\pi\pi^*$. Even though the computed spectrum is much more complicated than previously expected, it has very good agreement with the experimental data of this tautomer.

We performed a second analysis of the absorption spectrum of the keto N1H tautomer to confirm the results obtained by the classification of transitions previously made. We calculated in intervals of 0.2 eV the percentage of absorptions assigned to each type of electronic transition considered for the keto N1H form, the $\pi\pi^*$, $n_{\text{O}}\pi^*$, and $n_{\text{N}}\pi^*$. Additionally, we computed the population of the O and N atoms with lone electron pairs, namely the O2 and N3 atoms, by means of the LoProp partitioning scheme.⁷⁴ Figure 6 shows a graphical representation of both data for comparison. The charge variation on the N atom correlates well with the regions of the spectrum with large contributions of $n_{\text{N}}\pi^*$ transitions, the region between 5 and 6 eV, and the $n_{\text{N}}\pi^*$ transition at 6.9 eV. Moreover, the agreement between the charge gains of the O atom and the assignments of $n_{\text{O}}\pi^*$ transitions performed is also good. The larger charge gains for O are found primarily in the 5–6 eV region and at energies higher than 9 eV, where the $n_{\text{O}}\pi^*$ bands have been identified. The nonzero charge gain for O in some regions of the spectrum dominated by $\pi\pi^*$ transitions is due to the π orbital of the O atom, which usually forms a MO localized on the carbonyl group. Therefore, the $\pi\pi^*$ transitions involving that π MO will produce significant changes in the charge of the O atom. This is not the case for the N atom, which will delocalize the charge to some part

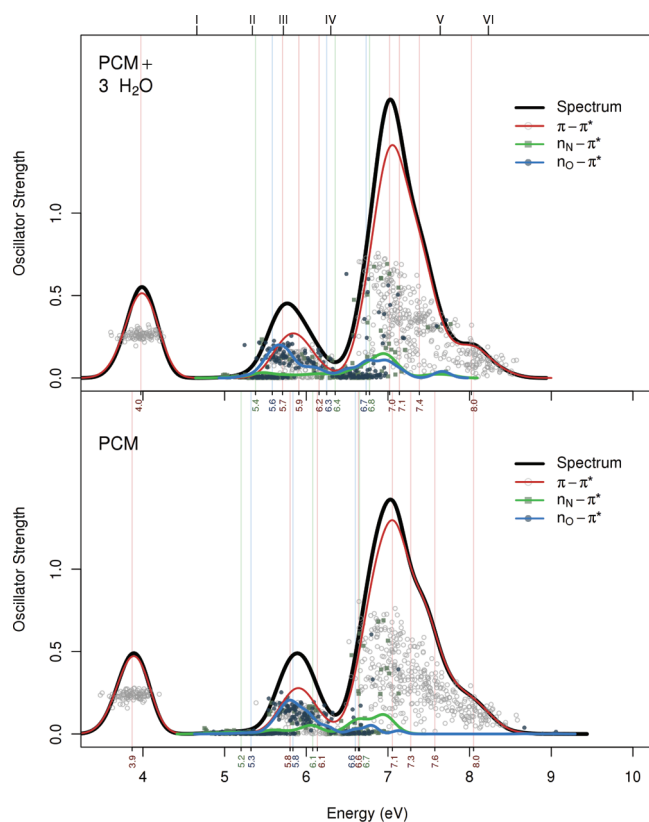


Figure 7. Computed absorption spectra of cytosine keto N3H tautomer. Result from 100 conformers using a PCM solvent model (bottom) and, additionally, with three explicit water molecules (top). Density estimation with Gaussian kernel functions ($bw = 0.12$ eV). Roman numbers on top mark the experimentally observed features.¹⁶

of the pyrimidine ring. This comparison shows that the breakdown performed on the absorption spectra is valid.

Keto N3H Tautomer. Figure 7 shows the computed absorption spectra of the cytosine keto N3H tautomer in water. This tautomer has one fewer $n_{O}\pi^{*}$ transition compared with the keto N1H. We have identified eight $\pi\pi^{*}$, three $n_{O}\pi^{*}$, and three $n_{N}\pi^{*}$ transitions in total. The number of these transitions is independent of the solvent model employed. The inclusion of explicit water molecules in the calculations of this tautomer does not affect much its electronic structure. The $n\pi^{*}$ transitions are slightly blue-shifted by ~ 0.3 eV, while the $\pi\pi^{*}$ transitions are in general lowered by ~ 0.2 eV. Qualitatively, the main effect of treating the hydrogen bonds is a larger differentiation between the different types of transitions that form the second band of the spectrum and a small shrinking of the main absorption band. However, we do not observe a high increase in intensity of any $n\pi^{*}$ transition, as was the case for the keto N1H form; this is ascribed to the small amount of $n\pi^{*}$ excitation in the main $\pi\pi^{*}$ band region.

The first band on the spectrum of the keto N3H tautomer is formed solely by transitions of the $\pi\pi^{*}$ type. The second observed band shows a more complex structure, composed primarily by one $n_{O}\pi^{*}$ with relatively high intensity, mixed with two $\pi\pi^{*}$ transitions. On the red edge of the second band, there is a $n_{N}\pi^{*}$ transition that does not contribute to the absorption. Similarly, in the optically inactive region between the second and third band, there are dark transitions of all types. The high absorption

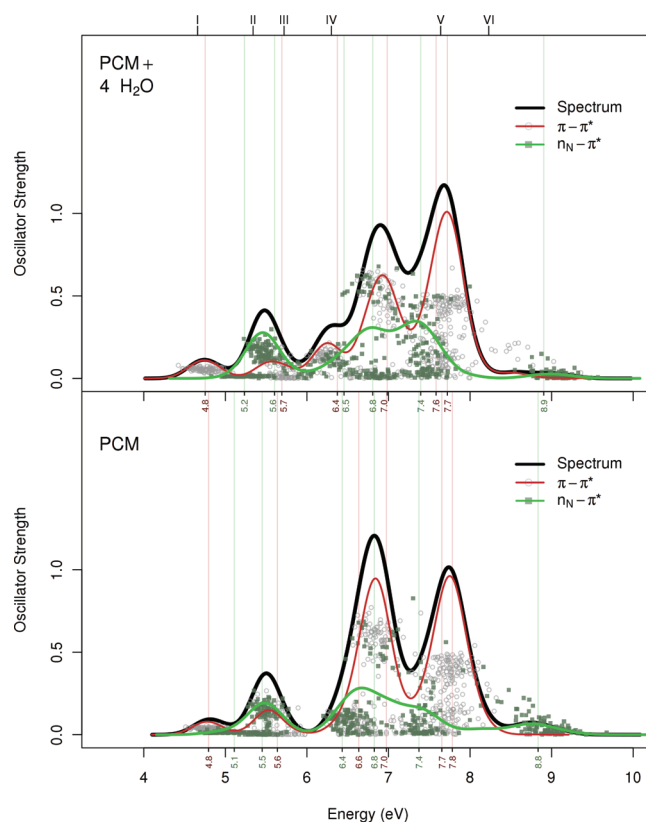


Figure 8. Computed absorption spectra of the cytosine enol tautomer. Result from 100 conformers using a PCM solvent model (bottom) and, additionally, with four explicit water molecules (top). Density estimation with Gaussian kernel functions ($bw = 0.14$ eV). Roman numbers on top mark the experimentally observed features.¹⁶

band of this spectrum has a maximum plus two shoulders at higher energies of decreasing intensity; all of these features are caused by intense $\pi\pi^{*}$ transitions.

Enol Tautomer. Figure 8 shows the computed absorption spectra of the cytosine enol tautomer in water. The spectral analysis of this tautomeric form is simpler compared to the keto and imino forms. There are no $n_{O}\pi^{*}$ transitions, and only one type of $n_{N}\pi^{*}$ can be considered because the lone pairs of the two N atoms are indistinguishable. Accordingly, six $\pi\pi^{*}$ and six $n_{N}\pi^{*}$ transitions are identified along all enol conformers. The inclusion of explicit water molecules produces small blue shifts on some $n_{N}\pi^{*}$ transitions and small red shifts of some $\pi\pi^{*}$ transitions, in line with the results of the previous tautomers. The largest effect is found in the third $\pi\pi^{*}$ transition at the beginning of the high absorption bands which forms a new visible band at 6.4 eV. Moreover, the $n_{N}\pi^{*}$ transitions located in regions of the spectrum with a large number of intense $\pi\pi^{*}$ transitions, like the second band at 5.5 eV and the two high absorption bands found between 6.5 and 8.0 eV, experience a significant gain of intensity upon inclusion of the water molecules in the CASPT2 calculation. Similarly to the other tautomers, this phenomenon arises from an increase of the π character of these transitions, increasing the overlap between the ground state and the involved excited state.

The peaks of absorption of the enol tautomer spectrum are well-defined and have contributions from few transitions, easing their assignments. The first and fifth absorption bands,

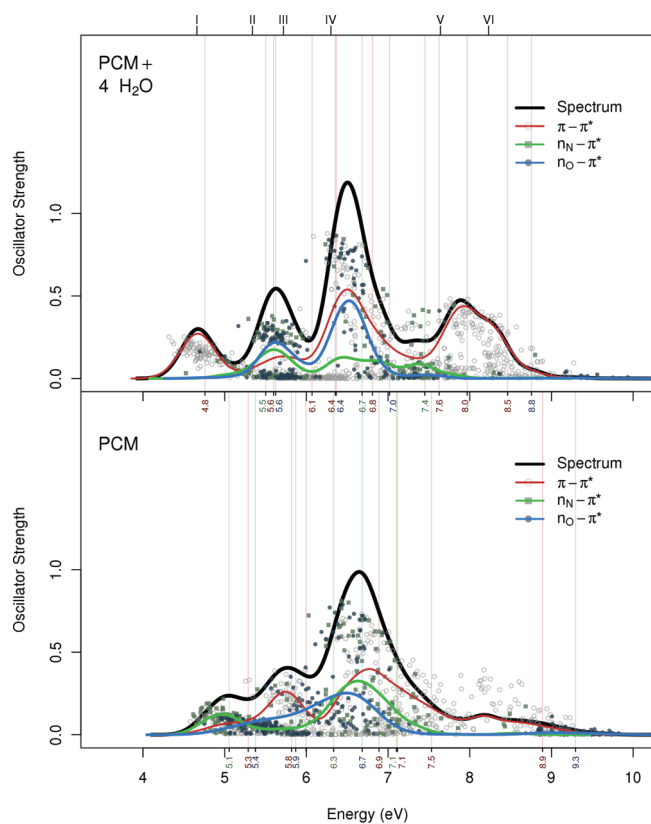


Figure 9. Computed absorption spectra of cytosine imino tautomer. Result from 100 conformers using a PCM solvent model (bottom) and, additionally, with four explicit water molecules (top). Density estimation with Gaussian kernel functions ($bw = 0.14$ eV). Roman numbers on top mark the experimentally observed features.¹⁶

at 4.8 and 7.7 eV, respectively, can be assigned to pure $\pi\pi^*$ bands. Instead, the second band is caused primarily by $n_N\pi^*$ excitations. The complexity of this spectrum increases for the third and fourth bands, at 6.4 and 6.9 eV, which appear as a mix with similar weights of both types of transition. Furthermore, there are some features which are masked in the spectrum. Located at 7.3 eV, in the valley between the two high absorption peaks of the fourth and fifth bands, there is a $n_N\pi^*$ band of lower intensity but that is optically active. The rest of transitions do not contribute to absorption, like the excitations found at ~ 9 eV.

Imino Tautomer. Figure 9 shows the computed absorption spectra of the cytosine imino tautomer in water. We have identified eight $\pi\pi^*$, three $n_O\pi^*$, and three $n_N\pi^*$ transitions along all conformers. This tautomer shows the largest change upon the addition of explicit water molecules in the solvent model. The basic morphology of the spectra is formed by three peaks of increasing intensity, which are present in both solvent models. However, the inclusion of water molecules into the imino tautomer forms a fourth band in the higher energy region of the spectra. This band appears by the concentration of $\pi\pi^*$ transitions that otherwise are found as a disperse cloud in the PCM spectrum. There is a relatively large blue shift of a $\pi\pi^*$ excitation from 7.1 eV in PCM to 8.0 eV including four waters. Moreover, there is also a red shift of another intense $\pi\pi^*$ transition from 8.9 to 8.5 eV. These intense absorptions are the main peak and shoulder of the fourth band. Additionally, the inclusion of water molecules to the imino form has two effects

that we already observed in the other tautomers. The first one is the increase in absorption of some $n_O\pi^*$ excitations located in the most intense band, due to an increase of their π character that leads to a larger overlap between the involved states. The second one is a higher differentiation between the transitions that contribute to each absorption band, which in this case is observed along all of the spectrum.

The first band on the absorption spectrum of the imino tautomer is a pure $\pi\pi^*$ band. However, the source of these excitations is not the π MO on the carbonyl group as for the two keto tautomers. Instead, they involve a π MO mainly localized on the amine group. These transitions are identified as $n_N\pi^*$ excitations in the spectrum without explicit water molecules. The hydrogen bonds formed with surrounding solvent molecules stabilize the lone pair in the N atom, increasing the $\pi\pi^*$ character of the lowest electronic excitations and pushing the first $n_N\pi^*$ transitions into the second absorption band. This band is formed by a mix of three transitions, each one of a different type, that contribute in the same proportion to the total absorption. Subsequently, an optically almost inactive $\pi\pi^*$ transition is identified at 6.1 eV. The third band is the most intense feature in the imino tautomer spectrum. One $\pi\pi^*$ and one $n_O\pi^*$ transition are the main excitations behind this band. Additionally, there is a weak $\pi\pi^*$ transition that produces a shoulder on the end of the third band. There are two dark transitions masked by the intense absorption, a $n_N\pi^*$ at 6.7 eV and a $n_O\pi^*$ at 7.0 eV. Furthermore, a small peak appears in the valley between the third and fourth bands caused by a $n_N\pi^*$ transition at 7.4 eV. The high end of the spectrum shows a low absorption produced by a $n_O\pi^*$ transition.

4. CONCLUSIONS

We have successfully combined the Car–Parrinello MD method with CASPT2 calculations to describe the absorption spectrum of four cytosine tautomers (Figure 1). We explored the conformational space of these tautomeric forms in water, extracted a representative sample of 100 conformers for each tautomer and studied their electronic structure with an efficient treatment of electron correlation. The wave function based calculations included the solvent effects through a PCM for water and some additional explicit water molecules. From the resulting data set, we built the spectrum of each tautomer and analyzed their underlying structure.

The computed absorption spectrum of the keto N1H tautomer has a very good agreement with previous experimental data (Figure 5). The combination of Car–Parrinello MD and CASPT2 generates a theoretical spectrum of cytosine that reproduces not only the energetics but also the bandwidths and intensities of the spectral features. The analysis of the calculated excitations makes it possible to identify them by their characteristics and describe the different individual contributions to the spectrum. The six observed transitions in the experimental spectrum are formed by a total of 15 transitions, which are grouped to give rise to the six experimental absorptions. The difference in energy between the calculated and the experimental bands is very low, on the order of 0.1 eV. A comparison with the LoProp population of the O and N atoms with lone electron pairs proved the classification of the different transition contributing to the spectrum to be correct (Figure 6). Additionally, following the same procedure performed on the keto N1H tautomer, we computed and analyzed the absorption spectra of the other three

tautomers. All of them show similar characteristics to the keto N1H tautomer but with significant differences on their spectrum composition.

The results obtained confirm the complexity of studying the cytosine molecule. Even though it is a relatively small molecule, its flexibility and tautomerization phenomena impose the use of dynamical techniques to cover vibro-rotational degrees of freedom that are not accessible from static approaches. Moreover, the increased complexity also affects the electronic structure and its analysis. Each tautomer shows different electronic properties which require an individualized study of them. The absorption spectra are formed by various transition types, which we classified in $n_{\text{O}}\pi^*$, $n_{\text{N}}\pi^*$, and $\pi\pi^*$ excitations. These archetypical transitions can mix due to the conformational distortions that break the planar symmetry of the ideal cytosine molecule.

The effect on the absorption spectrum of the hydrogen bonds between cytosine and solvent molecules is to shift both $n\pi^*$ and $\pi\pi^*$ transitions. However, there is not a clear trend on the change produced in the absorption energies. The $n\pi^*$ transitions are usually blue-shifted, but the effect on $\pi\pi^*$ transitions is very variable. We observe that the hydrogen bonds combined with the structural distortions of the conformer can favor even more the mix between excitations involving a nonbonding MO with those involving a π MO. Some $n\pi^*$ excitations experience a large increase of intensity due to their increased π character. This behavior shows that once the conformational distortions are incorporated into the theoretical model, it is no longer possible to get pure transitions of one single type. Thus, there is no simple description of the electronic structure of cytosine, even though its molecular structure and absorption spectra can look simple.

AUTHOR INFORMATION

Corresponding Author

*E-mail: alex.domingo@urv.cat; coen.degraaf@urv.cat.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

Financial support has been provided by the Spanish Ministry of Science and Innovation (Projects CTQ2008-06644-C02-01 and CTQ2008-06549-C02-01), the Generalitat de Catalunya (Project 2009SGR462 and *Xarxa d'R+D+I en Química Teòrica i Computacional, XRQTC*), and COST Action CODECS, CM1002.

REFERENCES

- (1) Middleton, C. T.; de La Harpe, K.; Su, C.; Law, Y. K.; Crespo-Hernández, C. E.; Kohler, B. *Annu. Rev. Phys. Chem.* **2009**, *60*, 217–239. PMID: 19012538.
- (2) Crespo-Hernández, C. E.; Cohen, B.; Hare, P. M.; Kohler, B. *Chem. Rev.* **2004**, *104*, 1977–2020. PMID: 15080719
- (3) Merchán, M.; Serrano-Andrés, L. *J. Am. Chem. Soc.* **2003**, *125*, 8108–8109. PMID: 12837073.
- (4) Serrano-Andrés, L.; Merchán, M. *THEOCHEM* **2005**, 729, 99–108. Proceedings of the 30th International Congress of Theoretical Chemists of Latin Expression, The 30th International Congress of Theoretical Chemists of Latin Expression.
- (5) Merchán, M.; González-Luque, R.; Climent, T.; Serrano-Andrés, L.; Rodríguez, E.; Reguero, M.; Peláez, D. *J. Phys. Chem. B* **2006**, *110*, 26471–26476. PMID: 17181307.
- (6) González-Vázquez, J.; González, L. *ChemPhysChem* **2010**, *11*, 3617–3624.
- (7) Barbatti, M.; Aquino, A. J. A.; Szymczak, J. J.; Nachtigallova, D.; Lischka, H. *Phys. Chem. Chem. Phys.* **2011**, *13*, 6145–6155.
- (8) Merchán, M.; Serrano-Andrés, L.; Robb, M. A.; Blancafort, L. *J. Am. Chem. Soc.* **2005**, *127*, 1820–1825. PMID: 15701017.
- (9) Blancafort, L. *Photochem. Photobiol.* **2007**, *83*, 603–610.
- (10) Kosma, K.; Schröter, C.; Samoylova, E.; Hertel, I. V.; Schultz, T. *J. Am. Chem. Soc.* **2009**, *131*, 16939–16943, PMID: 19874018.
- (11) Clark, L. B.; Tinoco, I. *J. Am. Chem. Soc.* **1965**, *87*, 11–15.
- (12) Clark, L. B.; Peschel, G. G.; Tinoco, I. *J. Phys. Chem.* **1965**, *69*, 3615–3618.
- (13) Voelter, W.; Records, R.; Bunnenberg, E.; Djerassi, C. *J. Am. Chem. Soc.* **1968**, *90*, 6163–6170.
- (14) Lewis, T. P.; Eaton, W. A. *J. Am. Chem. Soc.* **1971**, *93*, 2054–2056.
- (15) Callis, P. R.; Simpson, W. *J. Am. Chem. Soc.* **1970**, *92*, 3593–3599.
- (16) Zaloudek, F.; Novros, J. S.; Clark, L. B. *J. Am. Chem. Soc.* **1985**, *107*, 7344–7351.
- (17) Matos, J. M. O.; Roos, B. O. *J. Am. Chem. Soc.* **1988**, *110*, 7664–7671.
- (18) Petke, J. D.; Maggiora, G. M.; Christoffersen, R. E. *J. Phys. Chem.* **1992**, *96*, 6992–7001.
- (19) Alyoubi, A. O.; Hilal, R. H. *Biophys. Chem.* **1995**, *55*, 231–237.
- (20) Hug, W.; Tinoco, I. *J. Am. Chem. Soc.* **1973**, *95*, 2803–2813.
- (21) Hug, W.; Tinoco, I. *J. Am. Chem. Soc.* **1974**, *96*, 665–673.
- (22) Fülischer, M. P.; Roos, B. O. *J. Am. Chem. Soc.* **1995**, *117*, 2089–2095.
- (23) Billingham, B. E.; Oladepo, S. A.; Loppnow, G. R. *J. Phys. Chem. B* **2009**, *113*, 7392–7397. PMID: 19438283.
- (24) Tomić, K.; Tatchen, J.; Marian, C. M. *J. Phys. Chem. A* **2005**, *109*, 8410–8418. PMID: 16834234.
- (25) Dreyfus, M.; Bensaude, O.; Dodin, G.; Dubois, J. E. *J. Am. Chem. Soc.* **1976**, *98*, 6338–6349.
- (26) Szczesniak, M.; Szczepaniak, K.; Kwiatkowski, J. S.; KuBulat, K.; Person, W. B. *J. Am. Chem. Soc.* **1988**, *110*, 8319–8330.
- (27) Brown, R. D.; Godfrey, P. D.; McNaughton, D.; Pierlot, A. P. *J. Am. Chem. Soc.* **1989**, *111*, 2308–2310.
- (28) Feyer, V.; Plekan, O.; Richter, R.; Coreno, M.; de Simone, M.; Prince, K. C.; Trofimov, A. B.; Zaytseva, I. L.; Schirmer, J. *J. Phys. Chem. A* **2010**, *114*, 10270–10276.
- (29) Bazzo, G.; Tarczay, G.; Fogarasi, G.; Szalay, P. G. *Phys. Chem. Chem. Phys.* **2011**, *13*, 6799–6807.
- (30) Pedone, A.; Biczysko, M.; Barone, V. *ChemPhysChem* **2010**, *11*, 1812–1832.
- (31) Tomasi, J.; Persico, M. *Chem. Rev.* **1994**, *94*, 2027–2094.
- (32) Colominas, C.; Luque, F. J.; Orozco, M. *J. Am. Chem. Soc.* **1996**, *118*, 6811–6821.
- (33) Sambrano, J. R.; de Souza, A. R.; Queralto, J. J.; Andrés, J. *Chem. Phys. Lett.* **2000**, *317*, 437–443.
- (34) Mercier, Y.; Santoro, F.; Reguero, M.; Improta, R. *J. Phys. Chem. B* **2008**, *112*, 10769–10772. PMID: 18700794.
- (35) Improta, R.; Barone, V. *THEOCHEM* **2009**, 914, 87–93. Time-dependent density-functional theory for molecules and molecular solids.
- (36) Shukla, M. K.; Leszczynski, J. *J. Phys. Chem. A* **2002**, *106*, 11338–11346.
- (37) Broo, A.; Holmén, A. *J. Phys. Chem. A* **1997**, *101*, 3589–3600.
- (38) Kowalski, K.; Valiev, M. *J. Phys. Chem. A* **2008**, *112*, 5538–5541. PMID: 18505240.
- (39) Brancato, G.; Rega, N.; Barone, V. *Chem. Phys. Lett.* **2010**, *500*, 104–110.
- (40) Watson, J. D.; Crick, F. H. C. *Nature* **1953**, *171*, 737–738.
- (41) Olaso-González, G.; Roca-Sanjuán, D.; Serrano-Andrés, L.; Merchán, M. *J. Chem. Phys.* **2006**, *125*, 231102.
- (42) Kostko, O.; Bravaya, K.; Krylov, A.; Ahmed, M. *Phys. Chem. Chem. Phys.* **2010**, *12*, 2860–2872.
- (43) Barbatti, M. *Phys. Chem. Chem. Phys.* **2011**, *13*, 4686–4692.
- (44) Isayev, O.; Furmanchuk, A.; Shishkin, O. V.; Gorb, L.; Leszczynski, J. *J. Phys. Chem. B* **2007**, *111*, 3476–3480. PMID: 17388492.

- (45) Schyman, P.; Laaksonen, A.; Hugosson, H. W. *Chem. Phys. Lett.* **2008**, *462*, 289–294.
- (46) Lawson Daku, L. M.; Linares, J.; Boillot, M.-L. *Phys. Chem. Chem. Phys.* **2010**, *12*, 6107–6123.
- (47) Manzoni, V.; Lyra, M. L.; Gester, R. M.; Coutinho, K.; Canuto, S. *Phys. Chem. Chem. Phys.* **2010**, *12*, 14023–14033.
- (48) Barone, V.; Bloino, J.; Monti, S.; Pedone, A.; Prampolini, G. *Phys. Chem. Chem. Phys.* **2010**, *12*, 10550–10561.
- (49) Barone, V.; Bloino, J.; Monti, S.; Pedone, A.; Prampolini, G. *Phys. Chem. Chem. Phys.* **2011**, *13*, 2160–2166.
- (50) Valiev, M.; Kowalski, K. J. *Chem. Phys.* **2006**, *125*, 211101.
- (51) Barbatti, M.; Aquino, A. J. A.; Lischka, H. *Phys. Chem. Chem. Phys.* **2010**, *12*, 4959–4967.
- (52) Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471–2474.
- (53) Andersson, K.; Malmqvist, P.-Å.; Roos, B. O. *J. Chem. Phys.* **1992**, *96*, 1218–1226.
- (54) Roos, B. O.; Taylor, P. R.; Siegbahn, P. E. M. *Chem. Phys.* **1980**, *48*, 157–173.
- (55) Cossi, M.; Rega, N.; Scalmani, G.; Barone, V. *J. Chem. Phys.* **2001**, *114*, 5691–5701.
- (56) Marx, D.; Hutter, J. *Ab Initio Molecular Dynamics*; Cambridge University Press: Cambridge, U.K., 2009.
- (57) Troullier, N.; Martins, J. L. *Phys. Rev. B* **1991**, *43*, 1993–2006.
- (58) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.
- (59) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.
- (60) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695–1697.
- (61) Nosé, S. *J. Chem. Phys.* **1984**, *81*, 511–519.
- (62) Grossman, J. C.; Schwegler, E.; Draeger, E. W.; Gygi, F.; Galli, G. *J. Chem. Phys.* **2004**, *120*, 300–311.
- (63) CPMD; IBM Corp.: Endicott, NY, 1990; MPI für Festkörperforschung: Stuttgart, Germany, 1997.
- (64) Cremer, D.; Pople, J. A. *J. Am. Chem. Soc.* **1975**, *97*, 1354–1358.
- (65) Boeyens, J. C. A. *J. Chem. Crystallogr.* **1978**, *8*, 317–320.
- (66) Spek, A. L. *Acta Crystallogr., Sect. D* **2009**, *65*, 148–155.
- (67) Roos, B. O.; Lindh, R.; Malmqvist, P.-Å.; Veryazov, V.; Widmark, P.-O. *J. Phys. Chem. A* **2004**, *108*, 2851–2858.
- (68) Forsberg, N.; Malmqvist, P.-Å. *Chem. Phys. Lett.* **1997**, *274*, 196–204.
- (69) Malmqvist, P.-Å.; Roos, B. O. *Chem. Phys. Lett.* **1989**, *155*, 189–194.
- (70) 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. Proceedings of the Symposium on Software Development for Process and Materials Design.
- (71) Aquilante, F.; De Vico, L.; Ferré, N.; Ghigo, G.; Malmqvist, P.-Å.; Neogrady, P.; Pedersen, T. B.; Pitoňák, M.; Reiher, M.; Roos, B. O.; Serrano-Andrés, L.; Urban, M.; Veryazov, V.; Lindh, R. *J. Comput. Chem.* **2010**, *31*, 224–247.
- (72) Team, R. D. C. R. *A Language and Environment for Statistical Computing*. 2009. ISBN 3-900051-07-0.
- (73) Kessler, H. *Angew. Chem., Int. Ed.* **1970**, *9*, 219–235.
- (74) Gagliardi, L.; Lindh, R.; Karlström, G. *J. Chem. Phys.* **2004**, *121*, 4494–4500.

Tuned Range-Separated Time-Dependent Density Functional Theory Applied to Optical Rotation

Monika Srebro^{†,‡} and Jochen Autschbach^{*,†}

[†]Department of Chemistry, University at Buffalo, State University of New York, Buffalo, New York 14260-3000, United States

[‡]Department of Theoretical Chemistry, Faculty of Chemistry, Jagiellonian University, R. Ingardena 3, 30-060 Krakow, Poland

 Supporting Information

ABSTRACT: For range-separated hybrid density functionals, the consequences of using system-specific range-separation parameters (γ) in calculations of optical rotations (ORs) are investigated. Computed ORs at three wavelengths are reported for methyloxirane, norbornenone, β -pinene, [6]helicene, [7]helicene, and two derivatives of [6]helicene. The γ parameters are adjusted such that Kohn–Sham density functional calculations satisfy the condition $-\epsilon^{\text{HOMO}}(N) = \text{IP}$. For β -pinene, the behavior of the energy as a function of fractional total charge is also tested. For the test set of molecules, comparisons of ORs with available coupled-cluster and experimental data indicate that the γ “tuning” leads to improved results for β -pinene and the helicenes and does not do too much harm in other cases.

1. INTRODUCTION

Quantum-chemical methods based on Kohn–Sham density functional theory (DFT) and its time-dependent extension (TDDFT)^{1–4} have proven to be valuable tools in a broad variety of scientific fields including chemistry, biochemistry, physics, material sciences, spectroscopy, and catalysis.^{5–16} Despite their widespread popularity and great success in determination of properties for a wide variety of systems, several problems have been established in practical calculations that can be traced to the spurious electron self-repulsion present in commonly used approximations.^{17–23} In particular, many conventional generalized gradient approximations (GGA) and hybrid GGA exchange–correlation (XC) functionals fail in a qualitative and quantitative description of diffuse valence and Rydberg states as well as charge-transfer excitations due to incorrect asymptotic behavior and deficient long-range exchange. These issues affect computed molecular response properties, such as polarizabilities and optical rotations, to varying degrees ranging from insignificant to severe. Apart from early methods for treating self-interaction and recovering asymptotic behavior,^{24–27} a conceptually straightforward approach that has been proposed is the use of range-separated (long-range corrected, LC; Coulomb-attenuated method, CAM) functionals.^{28–31} As has been shown, such functionals offer remedies for origin problems.^{32–40}

In range-separated hybrid DFT the electron repulsion entering the exchange term of the Kohn–Sham energy functional is split into long-range and short-range parts by using for example the standard error function as in the Coulomb-attenuated method (CAM) of Yanai et al.³⁰

$$\frac{1}{r_{12}} = \frac{\alpha + \beta \operatorname{erf}(\gamma r_{12})}{r_{12}} + \frac{1 - [\alpha + \beta \operatorname{erf}(\gamma r_{12})]}{r_{12}} \quad (1)$$

Here, α and β are dimensionless parameters satisfying the relations $0 \leq \alpha + \beta \leq 1$, $0 \leq \alpha \leq 1$, and $0 \leq \beta \leq 1$. They

quantify the importance of the HF/DFT contribution in the short-range/long-range region. At r_{12} close to 0, the fraction of HF exchange is α , and its DFT counterpart is $1 - \alpha$. As r_{12} gets larger, the exchange is increasingly described by the HF expression rather than through DFT, approaching a fraction of $\alpha + \beta$ with r_{12} approaching ∞ . The range-separation parameter γ (in a_0^{-1} units) determines the balance of DFT to HF exchange at intermediate r_{12} , governing how rapidly the long-range limit is attained. A smaller/larger value results in slower/faster replacement of DFT exchange by its HF counterpart with an increase in interelectronic distances. For $\beta = 0$, the fraction of HF exchange is α over the whole range, which corresponds to conventional (global) hybrids. With $\alpha = 0$ and $\beta = 1$, the original LC approach of Iikura et al.²⁸ is reproduced, in which short-range exchange is represented purely by a local potential derived from the LDA or the GGA approximations. Fully long-range corrected functionals require $\alpha + \beta = 1$. The popular functional CAM-B3LYP does not fully switch to 100% HF but gives only 65% of exact exchange at large interelectronic distances with $\alpha = 0.19$ and $\beta = 0.46$ parameters determined through a fit to the atomization energies of a standard set of molecules.³⁰

As benchmark studies have shown, the performance of long-range corrected functionals is sensitive to the parametrization.^{34,35,41–44} In particular, a strong dependence of calculated ground- and excited-state properties on the range-separation parameter has been revealed. In most implementations, γ is adjusted in order to minimize average errors in equilibrium distances, atomization energies, barrier heights, ionization energies, and/or other ground-state properties of a test set of molecules and treated as a universal constant for subsequent computations. Typical γ ranges from 0.30 to 0.50 a_0^{-1} with recommended values of 0.33,²⁹ 0.40,^{28,33} 0.47,⁴⁵ and 0.50, depending on the functional.^{32,37} The optimal value differs,

Received: October 27, 2011

Published: November 30, 2011

however, for the properties of interest as well as for the specific systems being studied.^{41,43} A practical, physically motivated method for determining system-specific range-separation parameters has been recently suggested by Livshits and Baer in ref 37, utilized in refs 46–48. The suggested tuning procedure is based on the requirement that in exact Kohn–Sham theory, the negative of the energy of the highest occupied molecular orbital (HOMO) in the N electron system is equal to the ionization potential (IP) calculated as a ground-state energy difference

$$-\varepsilon^{\text{HOMO}}(N) = \text{IP} = E_{\text{gs}}(N-1) - E_{\text{gs}}(N) \quad (2)$$

Accordingly, γ can be adjusted to a given molecular system by minimizing

$$\Delta E(\gamma) = \varepsilon^{\text{HOMO}}(N; \gamma) + [E_{\text{gs}}(N-1; \gamma) - E_{\text{gs}}(N; \gamma)] \quad (3)$$

Above, the assumption is made that the same γ is appropriate for both the N and the $N-1$ electron systems. The tuned LC-(TD)DFT approach has been already successfully applied to several problems considered until recently as too difficult for DFT^{23,46–52} with the γ value varying substantially with the system under consideration. For example, for simple inorganic molecules, the optimal γ range is from $0.3 a_0^{-1}$ for Li_2 ; through $0.5 a_0^{-1}$ for CH_2O and NH_3 to $0.7 a_0^{-1}$ for HF , O_2 , and F_2 ; and $0.8 a_0^{-1}$ for P_2 .^{37,51} In all of these studies, the long-range corrected Baer, Neuhauser, and Livshits (BNL) functional^{31,37} was utilized. Herein, we adopt the γ tuning procedure in its original, simplest form³⁷ for other popular range-separated functionals and apply it in optical rotation calculations that were found to be notoriously challenging for TDDFT.

An alternative, related way of improving a functional in a system-specific way is to enforce the correct behavior of the energy E as a function of the electron number N ,^{36,53} which is also sometimes referred to as the straight line theorem:⁵⁴ The energy of an atom or molecule as a function of N changes linearly for N varying between two integers. The slope of $E(N)$ changes discontinuously as N passes through integers (derivative discontinuity). For instance, as N passes through the electron number of a neutral atom, the slope of $E(N)$ changes from $-\text{IP}$ (ionization potential) to $-\text{EA}$ (electron affinity). As an example, Vydrov et al. used the straight line criterion to assess the quality of the LC- ω PBE functional in comparison with standard GGAs and hybrid GGAs;³⁶ see also the Supporting Information (SI) accompanying this article (Figure S4). Similarly, Cohen et al.⁵³ and Tsuneda et al.⁵⁵ examined the energy for fractional charges with other range-separated functionals.

Optical rotation (OR) has been established as a valuable tool in the determination of the absolute configuration (AC) of chiral molecules. The protocol employs a comparison of a measured OR for a candidate stereoisomer with theoretical predictions for a known absolute configuration.^{56–62} To effectively utilize such a method, however, a robust and reliable approach to predicting optical rotations from first principles need to be applied. On the basis of studies by Stephens et al.,⁶³ the B3LYP hybrid functional with the aug-cc-pVDZ basis set based on geometry optimizations with the 6-31G(d) basis is often considered the standard protocol for OR calculations. The robustness and limitations of this level of theory have been explored in many benchmarks.^{60,64,65} Expected improvements for a description of Rydberg and charge-transfer states offered by range-separated functionals as compared to standard (global) hybrids and GGAs prompted us

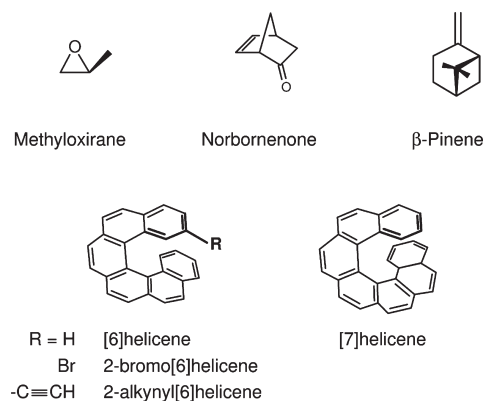


Figure 1. Molecular structures and absolute configurations of the systems studied herein.

recently to pursue a benchmark study (OR45) in order to examine the performance of such functionals in OR calculations for a diverse set of molecules ranging from small organic systems to organometallic complexes.⁶⁶ The intimate relation between the OR and the excitation spectrum implies that better performance of LC functionals should be observed. Our results, however, showed that, on average, the two range-separated functionals (LC-PBE0 and CAM-B3LYP) do not outperform their global hybrid counterparts for the OR45 test set.⁶⁶ This finding does not necessarily indicate a failure of the range-separation concept itself but may rather imply the need of ab initio-motivated molecule-specific reparametrization of the currently used LC functionals. The fact that a partial or full long-range correction was established to be beneficial in selected cases seems to corroborate this conclusion.⁶⁶ For instance, for norbornenone, the OR calculated with LC-PBE0/aug-cc-pVDZ was found to be close to coupled-cluster (CC) reference data and very different from the B3LYP result. The correct sign of the OR of β -pinene, a notoriously difficult case for computations, was also reproduced at this level of theory with the corresponding magnitude close to a gas-phase value interpolated from experimental data. Accordingly, taking into account the promising performance of tuned range-separated XC functionals in many types of computations,⁴⁹ and for Rydberg and charge-transfer excitations in particular, in the present study we examine such an approach to calculations of optical rotation for selected molecules of the OR45 benchmark differing in size and type of chromophores (Figure 1). The list of functionals studied herein includes LC-PBE and LC-BLYP, as well as the hybrids, LC-PBE0 and CAM-B3LYP (in its default and a modified parametrization). For literature references and parametrization details, see section 2. For the test systems, diffuse valence and Rydberg states (heterocycles, bicycles) and ‘charge-transfer-character’ excitations (helicenes), as postulated for linear and nonlinear polycyclic aromatic hydrocarbons,^{48,67} can be expected to contribute significantly to the OR, and thus a tuned TDDFT approach, ensuring their accurate description, should be especially advantageous.

In the following, we first provide additional technical details regarding the computations (section 2). Optical rotations obtained with tuned and nontuned functionals are reported along with experimental and CC reference data in section 3. A detailed analysis of the performance of fully long-range corrected functionals is provided, followed by a discussion of the behavior of

Coulomb-attenuated hybrids. A test of various XC functionals for a system with fractional electron numbers is also presented in this section. Finally, a brief summary and an outlook is given in section 4. The γ tuning appears to help with the ORs of β -pinene and the helicenes. In the other cases, when considering variations among the highest level calculations available in the literature, the results do not significantly deteriorate when a tuned γ parameter is used. On the basis of these findings, the tuning procedure can be cautiously recommended for applications to electronic optical activity.

2. COMPUTATIONAL DETAILS

Calculations were carried out for the molecules shown in Figure 1 in the absolute configurations as indicated, with the exception of 2-alkynyl[6]helicene. For this system, computations for the optical antipode were performed, and the corresponding optical rotations are reported here with the opposite sign of the ones calculated. The γ tuning procedure was performed for methyloxirane, norbornenone, β -pinene, and two helical systems, 2-alkynyl[6]helicene and [7]helicene. OR computations include also two other helicenes, [6]helicene and 2-bromo[6]helicene. Optimized structures for all of the systems studied here were taken from ref 66. Symmetry was not explicitly utilized in the calculations.

The computations were performed with a locally modified developer's version of the Northwest Computational Chemistry (NWChem) package^{68,69} using the augmented correlation-consistent Dunning basis set, aug-cc-pVDZ.^{70,71} Functionals examined in this work include long-range corrected variants of PBE^{72,73} and PBE0,^{74,75} labeled here as LC-PBE ($\gamma = 0.30 a_0^{-1}$, $\alpha = 0$, $\beta = 1$) and LC-PBE0⁴³ ($\gamma = 0.30 a_0^{-1}$, $\alpha = 0.25$, $\beta = 0.75$), and two parametrizations of the Coulomb-attenuated version of B3LYP:^{76–79} CAM-B3LYP in its original parametrization with $\alpha + \beta = 0.65$ ($\gamma = 0.33$, $\alpha = 0.19$, $\beta = 0.46$)³⁰ and a fully long-range corrected modification, LC-B3LYP, with $\alpha = 0.19$ and $\beta = 0.81$, $\alpha + \beta = 1.0$. Some computations utilized the LC variant of BLYP,^{76,77} LC-BLYP ($\gamma = 0.33 a_0^{-1}$, $\alpha = 0$, $\beta = 1$). In parentheses, the default parametrization of each functional as recommended in the NWChem manual is given. For comparison, OR calculations were carried out with B3LYP and Hartree–Fock (HF) as well. The ab initio system-specific determination of range-separation parameter γ was performed via minimizing the function given by eq 3. In single-point ground-state calculations of neutral and corresponding cation radical systems, an energy convergence threshold of 10^{-10} au was applied for all molecules with the exception of 2-alkynyl[6]helicene, for which 10^{-7} au was used.

For tests of the energy as a function of noninteger electron numbers, fractional orbital occupations and fractional total electron numbers were implemented in a developer's version of NWChem. Details will be provided elsewhere. The code was verified by a comparison of $E(N)$ for the carbon atom calculated with Hartree–Fock and various density functionals with reference data from the literature.³⁶ See the SI (Figure S4) for a plot of $E(N)$ for carbon.

Optical rotation (OR) calculations were carried out with a recently developed TDDFT linear response module (“AOResponse”) implemented in NWChem^{80,81} and utilized the GIAO dipole length gauge to ensure origin invariance of isotropic ORs. The optical rotation parameters were computed at the sodium line wavelength $\lambda = 589.3$ nm ($\omega = 0.07732$ au), as well as $\lambda = 355$ nm ($\omega = 0.128$ au) and 633 nm (0.0720 au) to compare with available

OR data from gas-phase cavity ring-down polarimetry (CRDP) measurements^{82–84} and coupled-cluster data from the literature. The “xfine” integration grid was employed in numerical integrations of the XC potential and the XC response kernel. Convergence criteria were set to 10^{-10} au and 10^{-6} au for the SCF and coupled perturbed Kohn–Sham (CPKS) procedure, respectively. A default parameter of 10^{-5} au was used in the computations to remove linearly dependent basis function combinations.

Optical rotations are discussed herein in terms of molar rotations (MRs), $[\phi]$. The calculated molecular OR parameter β can be converted to the observable specific rotation (excluding local field corrections or concentrations effects) via

$$[\alpha] = 7200 \text{ deg} \frac{\omega^2 N_A}{c^2 M} \beta(\omega)$$

and further to molar rotation by

$$[\phi] = [\alpha] \frac{M}{100}$$

In the equations above, the “circular” frequency of the perturbing field ω is in units of s^{-1} ; $\beta(\omega)$ is in units of cm^4 . Further, N_A , c , and M are Avogadro's number, the speed of light (in cm s^{-1}), and the molecular weight (in g mol^{-1}), respectively. The units of $[\alpha]$ and $[\phi]$ are $\text{deg}/[\text{dm} (\text{g}/\text{cm}^{-3})]$ and $\text{deg cm}^2 \text{ dmol}^{-1}$, respectively. To put our results in perspective, where available we have included both experimental as well as coupled-cluster results reported in the literature, converting specific rotations given in each case to molar rotations. For brevity, range-separation parameter units of a_0^{-1} and molar rotation units of $\text{deg cm}^2 \text{ dmol}^{-1}$ are frequently dropped from here on.

3. RESULTS AND DISCUSSION

3.1. Fully Long-Range Corrected Functionals Applied to OR Calculations. The calculated ΔE as a function of γ according to eq 3 for methyloxirane, norbornenone, and β -pinene using the aug-cc-pVDZ basis set and the fully long-range corrected functionals LC-PBE, LC-PBE0, and LC-B3LYP are graphically presented in Figure 2. A γ range from 0.1 to $0.5 a_0^{-1}$ was considered in each case. For norbornenone, some calculations with LC-PBE0 and LC-B3LYP with γ close to 0.5 failed to reach SCF convergence for the radical cation, and the corresponding data points in Figure 2 have been omitted. The numerical values of $\Delta E(\gamma)$ are collected in Tables S1–S3 of the Supporting Information (SI).

$\Delta E(\gamma)$ changes considerably within the γ ranges shown in Figure 2. In each case, however, a particular value of γ can be found below $0.4 a_0^{-1}$, for which the condition $\Delta E(\gamma) = 0$ is satisfied. In the following, the values of the range-separation parameter rendering a very small (close to 0) $\Delta E(\gamma)$ and corresponding to the (ionization from the HOMO) minimum are denoted as the optimal values γ^* .

The system-specific range-separation parameters differ from the global ones (0.30 for LC-PBE and LC-PBE0, 0.33 for LC-B3LYP) to varying degrees, with deviations ranging from 0.02 to 0.11. For the LC hybrids, the γ^* values are always decreased compared to the universal γ parameters, which implies that the contribution of exact exchange at short-range in such cases is smaller than assumed on the basis of the universal γ . The change is especially pronounced for β -pinene (0.11 difference). In the case of LC-PBE, the γ tuning has mixed directions, with γ^* decreased with respect to the universal value for β -pinene

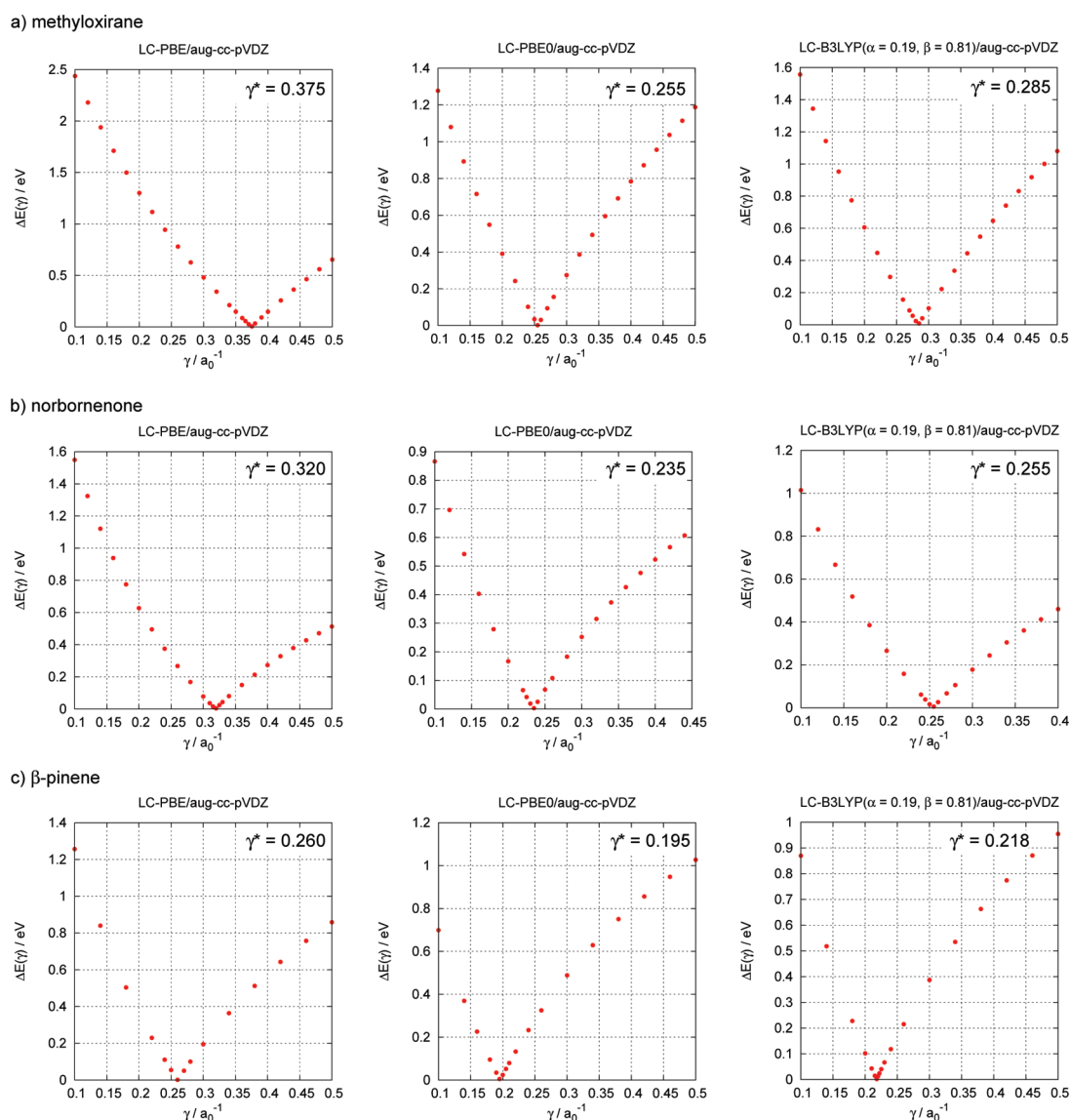


Figure 2. ΔE of eq 3 as a function of the range-separation parameter γ . Calculations with LC-PBE, LC-PBE0, and LC-B3LYP, for methyloxirane (top row), norbornenone (center row), and β -pinene (bottom row). The numerical γ^* values listed in the panels correspond to $\Delta E \approx 0$.

(by 0.04) but increased for norbornenone (by 0.02) and methyloxirane (by 0.08). In all cases, the smallest γ^* is determined for β -pinene and the largest for methyloxirane.

The corresponding molar rotations (MRs) at 355, 589, and 633 nm calculated using individually tuned γ^* and universal (empirically fitted) γ values are listed in Table 1 for LC-PBE and LC-PBE0 and in Table 2 for LC-B3LYP. For each molecule, experimental data and optical rotations obtained from approximate correlated wave function theories (CCSD and CC2) are provided as well.

There are many possible factors determining a reliable and accurate theoretical prediction of optical rotation within a given level of quantum-chemical methodology, among which the intrinsic quality of the electronic structure, conformational averaging, vibrational corrections, and solvent effects are of the highest importance. Since we have not included vibrational corrections and solvent effects in the calculations, it is more reasonable to assess the performance of the DFT methods by comparison with gas-phase single-point results from the best avail-

able wave function based ab initio methods, for instance, coupled-cluster theory, rather than experimental solution data. As reference data, we have chosen the recent CCSD results of ORs by Mach and Crawford⁸⁵ calculated with a modified velocity gauge (MVG)⁸⁶ and aug-cc-pVDZ basis set used in our study. In the case of norbornenone and β -pinene, due to the high computational cost, higher order CC data are not available in the literature. Comparisons with the results calculated with the double aug-cc-pVDZ basis set imply rather little basis set dependence of their ORs at the CCSD level. The situation is very different for methyloxirane. Noticeable differences between the CCSD values obtained with various basis sets as well as between optical rotations calculated with successive models of coupled-cluster theory of potentially increasing accuracy indicate that the chosen CCSD/aug-cc-pVDZ reference data are not converged in terms of basis sets and electron correlation effects.^{85,87,88} Accordingly, in the discussion, we have referred to CC3 data available in the literature as well. Some comparisons with gas-phase cavity ring-down polarimetry measurements are also made.

Table 1. Molar Rotations for Methyloxirane, Norbornenone, and β -Pinene Calculated with aug-cc-pVDZ and the LC-PBE and LC-PBE0 Functionals

	$[\phi]^{\text{exptl.a}}$	$[\phi]^{\text{CCSD}}/[\phi]^{\text{CC2b}}$	LC-PBE			LC-PBE0		
			$([\phi]\gamma = 0.47^d)$	γ^{*c}	$[\phi]\gamma^*$	$[\phi]\gamma = 0.30^c$	γ^{*c}	$[\phi]\gamma^*$
355 nm								
methyloxirane	4.35	-32.8/-43.3	-18.03 (-26.15)	0.375	-23.89	-21.60	0.255	-19.84
norbornenone		-4213.1/-6699.1	-8959 (-5598)	0.320	-8297	-6345	0.235	-7474
β -pinene	92.2	115.4/276.4	86.21 (0.952)	0.260	128.2	45.32	0.195	123.2
589.3 nm								
methyloxirane	-10.9	-17.5/-26.5	-12.05 (-11.90)	0.375	-12.25	-11.22	0.255	-11.23
norbornenone	-1239	-605.5/-880.6	-1044 (-815.2)	0.320	-1002	-892.9	0.235	-983.6
β -pinene	-31.5	1.0/34.7	-4.751 (-18.84)	0.260	1.725	-11.06	0.195	1.751
633 nm								
methyloxirane	-4.87	-15.4/-23.3	-10.67 (-10.42)	0.375	-10.77	-9.874	0.255	-9.908
norbornenone		-498.3/-721.4	-854.4 (-672.4)	0.320	-821.2	-735.2	0.235	-807.8
β -pinene	-6.35	-1.0/27.2	-5.587 (-17.12)	0.260	-0.310	-10.72	0.195	-0.234

^a 589.3 nm, solution data from refs 89–91; 355 and 633 nm, gas-phase data from ref 84. ^b CCSD and CC2/aug-cc-pVDZ/MVG data from ref 85. ^c Universal γ value as implemented in NWChem. ^d Universal γ value as implemented in Gaussian. ^e Optimal γ value as determined in this work.

Table 2. Molar Rotations for Methyloxirane, Norbornenone, and β -Pinene Calculated with aug-cc-pVDZ and the LC-B3LYP and CAM-B3LYP Functionals

	$[\phi]^{\text{exptl.a}}$	$[\phi]^{\text{CCSD}}/[\phi]^{\text{CC2b}}$	LC-B3LYP			CAM-B3LYP	
			$[\phi]\gamma = 0.33^c$	γ^{*d}	$[\phi]\gamma^*$	$[\phi]\gamma = 0.33^c$	$[\phi]\gamma^*$
355 nm							
methyloxirane	4.35	-32.8/-43.3	-16.82	0.285	-13.54	-13.13	-9.940
norbornenone	-5315.1	-4213.1/-6699.1	-6089	0.255	-7418	-7754	-8937
β -pinene	92.2	115.4/276.4	70.55	0.218	173.7	161.3	240.3
589.3 nm							
methyloxirane	-10.9	-17.5/-26.5	-10.61	0.285	-10.40	-10.83	-10.55
norbornenone	-1239	-605.5/-880.6	-860.8	0.255	-891.5	-999.5	-1080
β -pinene	-31.5	1.0/34.7	-7.150	0.218	9.040	8.400	19.86
633 nm							
methyloxirane	-4.87	-15.4/-23.3	-9.398	0.285	-9.257	-9.646	-9.440
norbornenone	-520.2	-498.3/-721.4	-708.9	0.255	-791.4	-820.0	-883.5
β -pinene	-6.35	-1.0/27.2	-7.543	0.218	5.681	5.247	14.56

^a 589.3 nm, solution data from refs 89–91; 355 and 633 nm, gas-phase data from ref 84. ^b CCSD and CC2/aug-cc-pVDZ/MVG data from ref 85. ^c Universal γ value as implemented in NWChem. ^d Optimal γ value as determined in this work.

In the following discussion, unsigned relative deviations of computed data with respect to reference values (experimental or other theoretical data), i.e. $\Delta^r = |[\phi]_D^{\text{calcd}} - [\phi]_D^{\text{ref}}|/|[\phi]_D^{\text{ref}}|$, in percent, will be utilized.

Methyloxirane. Consider first the data obtained for a small rigid three-ring member of the test set, methyloxirane, which is among the most extensively studied systems for optical rotation. Recently, Mach and Crawford⁸⁵ reported CCSD/aug-cc-pVDZ/MVG MRs

of -32.8 , -17.5 , and -15.4 deg cm² dmol⁻¹ for 355, 589.3, and 633 nm, respectively. These results deviate noticeably from the experimental gas-phase data, especially at 355 nm, for which CCSD gives the opposite sign. HF/aug-cc-pVDZ/GIAO results determined by us (355, -14.58 ; 589.3, -7.534 ; 633 nm, -6.640) agree with the coupled-cluster data in terms of sign, although the significant drop in the magnitude can be observed due to the lack of electron correlation effects. The B3LYP functional appears to perform somewhat better than CCSD in delivering optical rotations closer to experimental results: 0.811, -9.950 , and -9.036 deg cm² dmol⁻¹, respectively. The problem of discrepancies between computed CC and gas-phase experimental data for methyloxirane is now well understood and attributed to zero-point vibrational corrections (ZPVCs).^{88,92–94} The seemingly good performance of B3LYP at 355 nm was shown to be due to a significant underestimation of the lowest (Rydberg) excitation energy leading to a fortuitous positive shift in the MR toward the experimental result.⁸⁷ Accordingly, taking into account the importance of Rydberg states for this molecule, it is expected that the correct asymptotic behavior of range-separated XC functionals may be especially beneficial here, improving the agreement between DFT and CC results.

According to the data collected in Tables 1 and 2, the three LC functionals in their standard parametrization perform reasonably well for methyloxirane when compared to the coupled-cluster data. The DFT values are located somewhere between Hartree–Fock and CC. The relative deviations from the CCSD data are similar at 589.3 and 633 nm and range from 31 to 39% with the following ordering of functionals in terms of delivering the best agreement with the reference data: LC-PBE > LC-PBE0 > LC-B3LYP. As can be seen, the better agreement is obtained with vanishing short-range HF exchange contributions. At 355 nm both PBE-based functionals still outperform the fully long-range corrected version of CAM-B3LYP, LC-B3LYP. However, the best agreement with CC is now obtained for LC-PBE0. The corresponding relative deviations Δ^r are LC-PBE = 45, LC-PBE0 = 34, and LC-B3LYP = 49%. For the two longer wavelengths, using the LC-PBE functional with a γ value of $0.47 a_0^{-1}$ as implemented in the Gaussian package⁹⁵ increases only slightly the relative deviations from CC, by 0.9–1.7%. However, at 355 nm, a significant decrease is observed, by 25%. Accordingly, it appears that at this wavelength the MR is especially sensitive to the functional parametrization and the short-range exact exchange contribution.

Using the first-principles tuned range-separation parameter γ^* changes the results to an insignificant degree at 589.3 and 633 nm. The γ tuning noticeably influences the optical rotations at 355 nm, as should be expected from the discussion in the previous paragraph. An increase from 0.30 to the tuned value of $0.38 a_0^{-1}$ for LC-PBE leads to a MR closer to the coupled-cluster data with Δ^r decreased to 27%. This is in line with the aforementioned improved performance of LC-PBE with $\gamma = 0.47$. In the case of the tuned LC hybrid functionals, the tuning produces an increase of the relative deviation to CC data by 5.4% for LC-PBE0 and by 10% for LC-B3LYP.

Assuming that zero-point vibrational corrections (ZPVC) calculated at one level of theory are to a good degree transferable to another level of OR calculations, we have used CCSD ZPVCs from ref 93, 20.05 and 3.74 deg cm² dmol⁻¹ for 355 and 633 nm, respectively, in conjunction with our equilibrium ORs discussed above. Resulting MRs ranging from -6.105 to 6.513 at 355 nm and from -7.029 to -5.516 at 633 nm are closer to experimental values than vibrationally corrected coupled-cluster data of -12.8

and -11.7 . Keeping in mind the incomplete convergence of the CCSD/aug-cc-pVDZ OR reference data in terms of basis sets and electron correlation effects, to get additional insight into the performance of the tuned TDDFT approach for this molecule, we have recalculated MRs at the LR-PBE level with the aug-cc-pVDZ, aug-cc-pVTZ,^{70,71} and d-aug-cc-pVDZ^{70,71,96} basis sets for C, O, and H, respectively, and compare them with available corresponding CC3 results of -13.5 and -10.3 deg cm² dmol⁻¹ at 355 and 589.3 nm.⁸⁸ Our MRs are -16.71 and $-11.30/-22.63$ and -11.55 for the nontuned/optimal γ parameter. There is better agreement between standard parametrizations of LC DFT and these CC3 data than what is discussed above for the CCSD/aug-cc-pVDZ reference values. For tuned LC-PBE, at 589.3 nm, we find $\Delta^r = 24\%$ with respect to CC3, but at 355 nm, $\Delta^r = 68\%$. Therefore, the assessment is inconclusive. Taking into account a strong dependence of optical rotation of methyloxirane on the CC level and the basis set quality, as well as structural parameters,⁸⁷ further studies appear to be in order to reliably assess the methyloxirane results for the tuned γ parameters.

Norbornenone. Norbornenone has a very large optical rotation at 589.3 nm, which was attributed to electronic coupling between the two π systems, C=C and C=O, present in this molecule.^{97,98} CCSD has been shown to strongly underestimate the experimental solution-phase MR.^{65,85,99} Likewise, we have recently shown that range-separated hybrids give MRs for norbornenone that are in some cases significantly below the solution-phase experimental values.⁶⁶ Nonhybrid DFT significantly overestimates the optical rotation.^{66,100,101} Vibrational corrections were shown to be relatively minor when compared to the equilibrium optical rotation.¹⁰²

We approach the discussion with the assumption that the available CCSD data are representative of gas-phase measurements. CCSD/aug-cc-pVDZ/MVG MRs of norbornenone calculated recently by Mach and Crawford⁸⁵ at 355, 589.3, and 633 nm are -4213 , -605.5 , and -498.3 deg cm² dmol⁻¹, respectively. They are comparable to HF data (GIAO) of -3389 , -645.6 , and -537.8 , which suggests that electron correlation, although not negligible, is not a major influence for the optical rotation of norbornenone. On the other hand, density functionals show a large variation in the calculated MRs, in particular as a function of HF exchange.⁶⁶ Thus, the main factor influencing the norbornenone OR is likely the exchange, or lack thereof, in approximate functionals. Since CCSD and HF yield MRs that are much closer to each other than the variations between pure GGAs and global hybrids, for instance, correlation effects offered by DFT are potentially of secondary importance for norbornenone, while exact exchange becomes of paramount relevance (see below for further discussion.)

As Tables 1 and 2 show, the standard NWChem parametrizations of asymptotically corrected LC-PBE, LC-PBE0, and LC-B3LYP functionals lead to MR values ranging from -6089 (45%) to -8959 (113%) at 355 nm, from -860.8 (42%) to -1044 (72%) at 589.3 nm, and from -708.9 (42%) to -854.4 deg cm² dmol⁻¹ (72%) at 633 nm (relative deviations from CC in parentheses). Although these functionals overestimate the MR magnitude, they yield better agreement with CCSD than the global hybrid B3LYP, which produces MRs of -12370 (193), -1291 (113), and -1051 (111%) at 355, 589.3, and 633 nm, respectively. In terms of agreement with CCSD, the functional performance is LC-B3LYP > LC-PBE0 > LC-PBE. The agreement becomes better as the fraction of the HF exchange in the functional at shorter interelectronic separations, determined by γ as well as

the α parameters, increases. This trend is further confirmed by LC-PBE calculations using the Gaussian value of $\gamma = 0.47$. In this case, the larger γ translates to a larger HF exchange contribution already at shorter interelectronic distances, which may compensate the lack of a fixed global HF contribution in this functional ($\alpha = 0$). The increased γ relative to the default parameters led to a significant decrease of Δ' , between 37 and 80%, at each wavelength. Accordingly, this parametrization of LC-PBE outperforms both LC hybrids.

The results calculated with the tuned functionals remain in line with our finding that the larger HF contribution at short-range improves the agreement with CCSD. In the case of both range-separated hybrids for which the tuned γ^* is smaller than the default value, the agreement with CCSD somewhat deteriorates. For LC-PBE, γ^* is slightly higher than the default value, and improved agreement with CCSD is obtained. As in the case of methyloxirane, for all functionals, the short wavelength optical rotation is more sensitive to the range-separation parameter value/short-range HF contribution in terms of variations in Δ' .

Although the optical rotations of norbornenone calculated at both tuned and nontuned LC-TDDFT levels remain significantly different from the CCSD results, it is worth noting that they match very well with results of the CC2 (second-order approximate CCSD) model.¹⁰³ This may imply that in terms of reproducing correlation effects the DFT methods considered here are closer in performance to CC2 than to CCSD. Using the optimized range-separation parameter generally leads to somewhat lessened agreement with CC2 for the LC-PBE0 and LC-B3LYP functionals and some improvements for LC-PBE. The sizable differences between CC2 and CCSD, however, also raise questions about the convergence of the wave function results with respect to the level of correlation. Going from HF to CC2 indicates that correlation plays a significant role in the optical rotation of norbornenone. The CCSD data are closer to HF again, reversing the trend from HF to CC2. One might infer from these trends that either CC2 artificially produces correlation effects that are not really present in the molecule or that lower order and higher order correlation effects cancel to a large degree. In the latter case, convergence might be difficult to achieve. When assessing the performance of the density functionals, it is important to keep in mind that, because the HF and CCSD reference data are reasonably close to each other, parameter changes that produce more HF exchange overall at the expense of correlation will give better agreement with the reference data. If there is a hidden role of correlation in the final result, the seemingly better performance obtained from such parameter tweaks might be for the wrong reason. We tentatively attribute the improvements from range-separated LC exchange, compared to pure DFT functionals and global hybrids, to the correct asymptotic behavior of the XC potentials.

During the course of this study, we noticed that the best agreement with CCSD is obtained with the LC-BLYP functional in its parametrization as implemented in the Gaussian package ($\gamma = 0.47$), yielding molar rotations of -5462 , -799.1 , and -659.1 deg cm² dmol⁻¹ at 355, 589.3, and 633 nm, respectively (slightly outperforming LC-PBE with the same γ). Tuning leads to $\gamma^* = 0.32 a_0^{-1}$. For the $\Delta E(\gamma)$ graph and the numerical data, see Figure S1 and Table S2 of the SI. The corresponding calculated MRs at 355, 589.3, and 633 nm are -8202 , -987.8 , and -809.1 , thus strongly increased in magnitude as compared to the results with the $\gamma = 0.47$. For comparison, the default NWChem parametrization gives results of -7774 , -960.2 , and -787.2 ,

which are bracketed by those obtained with higher and lower γ . The decrease in the MR magnitude with an increased HF contribution at shorter range remains in line with the findings for the other functionals.

β -Pinene. β -pinene is among the most problematic cases for OR calculations; most levels of theory fail to reproduce the sign of the experimentally observed optical rotation. Recent CCSD data by Mach and Crawford⁸⁵ (355, 115.4; 589.3, 1.0; and 633 nm, -1.0 deg cm² dmol⁻¹) agree reasonably well with CRDP measurements of 92.2 and -6.4 at 355 and 633 nm, as well as with an interpolated gas-phase value of -3.8 at 589.3 nm.⁶⁵ The large deviation between CCSD and the liquid-state MR at the sodium D line must be attributed to solution-phase experiment versus gas-phase calculation. HF theory strongly underestimates the MR at 355 nm (19.57) and overestimates it at long wavelengths (-16.73 and -15.49 deg cm² dmol⁻¹), leading also to the wrong sign at 589.3 nm. B3LYP produces MRs that are too high in magnitude (342.2, 34.76, 26.70) and fails to reproduce the CCSD sign for both of the long wavelengths.

The LC functionals in their default parametrizations perform somewhat better than HF and B3LYP, leading in each case to noticeably decreased relative deviations from the CCSD reference data. Again, an underestimation of the MR at 355 nm is observed, with the relative deviations from CCSD ranging from 25 to 61%. At 589.3 nm, none of the functionals considered here reproduce the sign and magnitude of the CCSD MR. At 633 nm, the signs agree with CCSD but, as in the case of the sodium D line, with a significant overestimation of the absolute value. The relative deviations range from 575 to 1206% and from 459 to 972% for 589.3 and 633 nm, respectively, owing to the small magnitude of the reference value. In terms of delivering the best agreement with the CCSD, the ordering of the functionals is LC-PBE > LC-B3LYP > LC-PBE0. For LC-PBE, using $\gamma = 0.47$ as a default in the Gaussian package leads to a significant deterioration of the calculated MR. Especially at 355 nm, the calculated MR of 0.952 deg cm² dmol⁻¹ is far from the expected range of values. However, the result is in line with the trend of increasing MR with decreasing γ as detailed in the next paragraph. Overall, most of the TDDFT calculation values produce reasonable agreement with experimental data (generally better than the agreement with CCSD), with Δ' ranging from 6.5 to 77%.

According to the data listed in Tables 1 and 2, the LC PBE-based (GGA and hybrid) functionals with the optimized range-separation parameter perform very similar in the β -pinene optical rotation calculations. A substantial improvement of the MRs is obtained from the γ tuning, with relative deviations from CCSD of 6.8 to 11 at 355, 72 to 75 at 589.3, and 69 to 77% at 633 nm. In each case, the CCSD OR sign is reproduced. Although the tuned LC-PBE and LC-PBE0 results are closer to the coupled-cluster MRs, they are further away from experimental results. The tuned parametrization of LC-B3LYP gives a deterioration rather than improvement, with the MR sign reproduced only for 355 and 589.3 nm and highly overestimated magnitudes.

Pristine and Substituted Helicenes. Recently, LC functionals have been shown to provide more reliable excitation energies for planar polyaromatic hydrocarbons and their helical isomers than conventionally used (TD)DFT approximations.^{48,67,104–106} It has been postulated that the improved performance is linked to a partial charge-transfer character of the $\pi \rightarrow \pi^*$ transitions in extended π chromophores.^{48,67} Likewise, it has been found

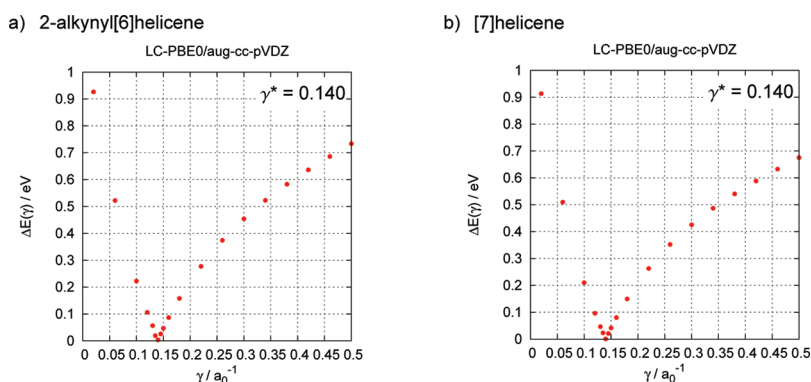


Figure 3. ΔE of eq 3 as a function of the range-separation parameter γ . LC-PBE0 calculations for 2-alkynyl[6]helicene (panel a) and [7]helicene (panel b). The printed γ^* values correspond to $\Delta E \approx 0$.

recently that range-separated hybrid functionals in their standard parametrizations tend to outperform global hybrids in optical rotation calculations of helicenes systems.⁶⁶ A counterexample is [7]helicene where a significant underestimation of the MR magnitude with respect to the solution-phase experiment was obtained, in particular for LC-PBE0. Taking into account promising results for excitation energies of oligoacene series and related hydrocarbons obtained with γ -tuned TDDFT,⁴⁸ it is therefore interesting to investigate whether system-specific γ parameters lead to improved optical rotations of helicenes and helicene derivatives.

Results of the tuning procedure performed for 2-alkynyl[6]helicene and [7]helicene with LC-PBE0 are graphically presented in Figure 3. The corresponding numerical data for $\Delta E(\gamma)$ are collected in Table S4 of the SI. The optimal γ parameter determined according to eq 3 is the same, $0.14 a_0^{-1}$, for both systems. This value nicely corresponds to results obtained for a series of oligoacenes $C_{2+4n}H_{4+2n}$ ($n = 1-6$) reported recently in refs 47 and 48, for which the tuned γ parameters decrease with system size, from 0.31 for benzene ($n = 1$) to 0.19 for hexacene ($n = 6$). Qualitatively, one may relate the decrease in the optimal γ to an increased range of delocalization in the π systems, making it beneficial to have short-range DFT exchange present in the functional at comparatively large interelectronic distances. The optimized range-separation parameter value is significantly decreased compared to the default parameter on the order of 0.3 typically used.

Calculated molar rotations (LC-PBE0/aug-cc-pVDZ) are collected in Table 3 along with solution-phase experimental data. The γ^* value determined for 2-alkynyl[6]helicene and [7]helicene was also applied in calculations of pristine [6]helicene and its bromo-substituted derivative. The following conclusions can be drawn on the basis of the presented data. (i) In the case of [7]helicene, using the optimized range-separation parameter in place of 0.30 leads to a significant improvement of the calculated MR toward the solution-phase experiment, with the relative deviation decreasing from 36 to 15%. For 2-alkynyl[6]helicene, in turn, the relative deviation increases from 19 to 56%. (ii) Optical rotations for [6]helicene and 2-bromo[6]helicene benefit from the optimized γ value, with a Δ' of merely 5 and 6%. (iii) The dependence of the helicenes OR on structural parameters was previously studied in ref 66. Using geometries optimized with Grimme's dispersion-corrected DFT (specifically: DFT-D3)¹¹¹ leads to a deterioration of the calculated optical rotations. Although

Table 3. Molar Rotations (in $\text{deg cm}^2 \text{ dmol}^{-1}$) for Helicenes and Helicene Derivatives Calculated with LC-PBE0/aug-cc-pVDZ

	$[\phi]^{\text{expt.}a}$	$[\phi]^{\text{calcd.}}$	
		$\gamma = 0.30^b$	$\gamma^* = 0.14^c$
2-alkynyl[6]helicene	−11042	−13183	−17245
[7]helicene	−23465	−15108	−19980
		−12048 ^d	−15818 ^d
			(−15709) ^d
[6]helicene	−11954	−9921.4	−12569
2-bromo[6]helicene	−14500	−11937	−15379

^a Solution data from refs 107–110. ^b Universal γ value, default parametrization. ^c Optimal γ as determined in this work for 2-alkynyl[6]helicene and [7]helicene. ^d Molar rotation for structure optimized with DFT-D3; in parentheses: molar rotation obtained with optimal γ of $0.143 a_0^{-1}$ determined for DFT-D3 geometry.

the tuned LC-PBE0 with $\gamma = 0.14$ subsequently improves the result for [7]helicene, the underestimation of the optical rotation with DFT-D3 geometries remains substantial (relative deviation of 33%). The tuning procedure performed for the DFT-D3 structure (Figure S2 and Table S4 of Supporting Information) leaves the optimal range parameter almost unchanged (0.143).

On the basis of computations on large carbon structures,^{111,112} the DFT-D structures are expected to significantly better represent the gas-phase helicene geometries. The effect of dispersion corrections on structural parameters of 2-bromo[6]helicene is detailed in the Supporting Information of ref 66. As was shown, dispersion corrections appear to substantially overestimate interactions between the aromatic rings at opposing ends of the helicene moieties when the optimized geometries are compared to the X-ray crystal structure. Accordingly, it seems likely that dispersion interactions are quenched in the crystal environment. The good agreement between nondispersion-corrected DFT and experimental geometries of organometallic helicene complexes^{113,114} corroborates this hypothesis. A similar situation is plausible also for the liquid phase, although further work on solvated systems would appear necessary to study these effects in more detail. We tentatively consider the structures optimized without dispersion corrections to be more suitable to reproduce solution-phase ORs. The solvent may also cause direct influences on the optical rotations which are not modeled. The UV–vis spectrum of

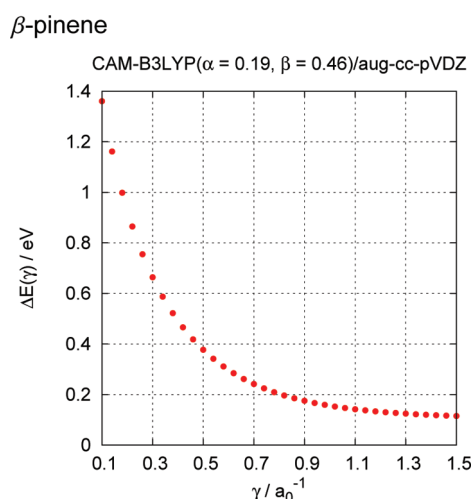


Figure 4. ΔE of eq 3 as a function of the range-separation parameter γ calculated with CAM-B3LYP for β -pinene.

these systems is dominated by valence excitations. It is therefore expected that condensed-phase effects are modest, compared to systems such as methyloxirane and the bicyclic cage structures investigated herein for which excitations involving diffuse orbitals are very important already at comparatively long wavelengths.

3.2. Coulomb-Attenuated Method CAM-B3LYP Applied to OR Calculations. The CAM-B3LYP functional utilizes $\alpha = 0.19$ and $\beta = 0.46$ in its original parametrization, minimizing errors for atomization energies for a test set of molecules.³⁰ Thus, by switching to only 65% of HF exchange at large interelectronic distances, the CAM-B3LYP XC potential V_{XC} does not afford the correct $-1/r$ behavior asymptotically. The different features of CAM-B3LYP as compared to LC functionals makes it an interesting case in the context of adjusting the range-separation parameter.

Tozer and Handy,²⁷ going back to arguments put forward by Perdew et al.,^{54,115} showed that, for continuum functionals, V_{XC} does not go to zero asymptotically. Rather, $V_{XC}(\infty) = IP + \varepsilon^{HOMO}$, where both values on the right-hand side correspond to calculations with the same functional. Tozer²¹ later showed that the charge-transfer failure of TDDFT, which is corrected by LC functionals, is intimately related to the integer discontinuity in the XC potential, $\Delta = V_{XC}^+ - V_{XC}^-$, which globally shifts the potential by a constant as the electron number for the system passes through an integer. The “+” and “−” indicate here a system with a slight excess and a slight deficiency of a fractional electron. “Pure” LDA and GGA density functionals do not exhibit this discontinuity. Tozer showed that in this case $\varepsilon^{HOMO} \approx -IP + \Delta/2$.²¹ Thus, the situation where $V_{XC}(\infty) = IP + \varepsilon^{HOMO} \neq 0$ is likely connected to the integer discontinuity problem where $\varepsilon^{HOMO} + IP \approx \Delta/2 \neq 0$. A pure DFT component in a functional that is not fully long-range corrected to enforce $V_{XC}(\infty) = 0$ may therefore prevent the condition of eq 3 to be fulfilled, except for the extreme case where $\gamma \rightarrow \infty$, which would completely suppress the short-range DFT exchange.

Figure 4 shows results for eq 3 for β -pinene calculated with CAM-B3LYP for different values of γ . Similar plots for methyloxirane and norbornenone can be found in Figure S3 of the SI. The corresponding numerical values $\Delta E(\gamma)$ are listed in Tables S1–S3 of the SI. Contrary to the fully long-range corrected

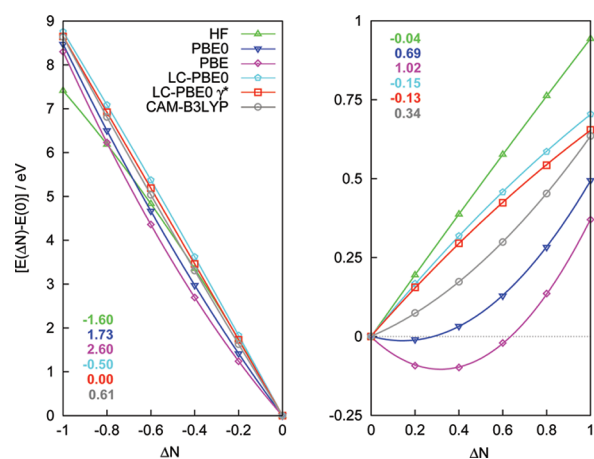


Figure 5. Ground-state energy (eV) of β -pinene as a function of deviations of fractional occupation number ΔN , $-1 \leq \Delta N \leq 0$ (left) and $0 \leq \Delta N \leq 1$ (right), calculated as the difference between actual and overall number of electrons in the neutral molecule. The energy is given relative to the energy of the neutral molecule ($\Delta N = 0$). Calculations were performed with the aug-cc-pVDZ basis set. The numerical values provide quantitative measures of curvature obtained from fitting quadratic functions to the data sets. For a single plot covering the full range $-1 \leq \Delta N \leq 1$ see Figure S5 of the SI.

functionals, no minimum is seen in Figure 4 even for a large range of γ values. Thus, the numerical data obtained for our examples support the qualitative arguments put forward in the previous paragraph, namely, that a functional that cannot establish $V_{XC}(\infty) = 0$ may not be “tunable” in the sense of eq 3.

For further tests with CAM-B3LYP, we have adopted the tuned γ^* for the LC-B3LYP functional. The MRs for CAM-B3LYP obtained with $\gamma = 0.33$ and with the γ^* are collected in Table 2. The smaller γ^* leads in most cases to worse agreement with CCSD optical rotations. The increase in relative deviations is the lowest for methyloxirane (by 1.3 to 9.7%), followed by norbornenone (13 to 28%) and β -pinene (by 68 to 1146%). For norbornenone, the CAM-B3LYP molar rotations are noticeably larger in magnitude as compared to LC-B3LYP. Accordingly, higher deviations from the CCSD and CC2 reference data are obtained. These results are in line with the increase of optical rotation magnitude when going from a fully long-range corrected LC-B3LYP of 100% to the standard global hybrid B3LYP with 20% exact exchange. The partially long-range corrected CAM-B3LYP (65% of HF exchange asymptotically) gives optical rotations between those calculated with LC-B3LYP and B3LYP. As far as β -pinene is concerned, the CAM-B3LYP parametrizations lead to MRs in least agreement with CCSD and fail to reproduce the sign at 633 nm. For all three wavelengths, a significant overestimation of the MRs can be observed. We note in passing that the original parametrization of CAM-B3LYP performs somewhat similar to the LC version with γ^* .

3.3. Calculations with Fractional Electron Numbers. The straight-line behavior of $E(N)$ was mentioned briefly in the Introduction. In this last section, we examine this behavior for β -pinene. Figure 5 shows calculated energies as a function of N around the electron number for the neutral molecule. With all density functionals as well as HF, the electron affinity is calculated to be negative. Correspondingly, the LUMO energy in the HF calculation is positive, indicating that the system will not bind an additional electron. The resulting derivative discontinuity in the $E(N)$ plot is large around $\Delta N = 0$, where the

slope changes formally from $-IP$ to $-EA$. A minor negative curvature in the excess electron section ($\Delta N > 0$) of the plot is seen for the standard parametrization of LC-PBE0, but overall a nearly optimal straight line behavior is obtained. The “tuned” version is slightly better but still displays some residual negative curvature in the electron-rich part of the plot ($\Delta N > 0$). The similarity between $E(N)$ obtained with the two parametrizations is surprising at first sight, given a significant change from the default $\gamma = 0.30$ to $\gamma^* = 0.20$ to satisfy eq 3. However, this is likely a consequence of the criterion adopted here, which predominantly affects the electron deficient side of the plot and, for the tuned version of LC-PBE0, indeed leads to a vanishing curvature for $\Delta N < 0$. Interestingly, for β -pinene, HF theory gives an essentially perfect straight line for $\Delta N > 0$ (much unlike the carbon atom, see the Figure S4 of the SI), albeit with a larger slope than LC-PBE0. Other functionals considered here yield a more (PBE, PBE0) or less (CAM-B3LYP) pronounced positive curvature for $E(N)$ which is typical of functionals with delocalization errors, as discussed in refs 17 and 116.

4. CONCLUSIONS

The success or failure of system-specific adjustments of the parameters in range-separated density functionals to satisfy eq 3 or, alternatively, the straight line theorem, is perhaps best judged by the principle *primum non nocere* (“first, do no harm”). This work has examined the consequences of selecting system-specific range-separation parameters in the context of calculating optical rotations for selected difficult cases. Optical rotation (OR) is a molecular mixed electric–magnetic dynamic linear response property that is known to be sensitive to approximations made in the electronic structure model. For small molecules, comparisons were made with available coupled-cluster data (CC2, CCSD, CC3). For the interesting helicene systems, reliable CC level calculations have unfortunately not yet been reported. The performance of the functionals for helicenes and helicene derivatives must be assessed with the help of experimental data under the assumption that for these systems solution-phase effects are minor.

For β -pinene and the helicenes, the system-specific adjustment of γ tends to produce improved ORs. For the helicenes, the optimized γ of 0.14 is significantly below commonly used values (typically on the order of 0.3) which can be attributed to the extended delocalized π systems. For norbornenone, the exchange component of the functional has a drastic influence on the calculated ORs. Our results obtained with LC functionals agree reasonably well with CC2 reference data, but they are too high in magnitude when compared to available CCSD optical rotations. For methyloxirane, adjusting γ does not lead to systematic improvements, but it also appears to do no particular damage. On the basis of this initial study, one may cautiously recommend the use of system-specific range-separation parameters for response calculations in the sense that it might either improve calculated ORs while at the same time ensuring some fundamental DFT requirements, or at least not do much harm. It remains to be seen if catastrophic failures upon γ tuning will be encountered. High values of γ appear to improve calculated ORs in selected cases, but such magnitudes of γ are difficult to justify on the basis of eq 3. It is worth emphasizing at this point that the tuning procedure in its original, simplest form as proposed by Livshits and Baer in ref 37 was applied herein. Modifications have been recently proposed^{46–48} in which, for example, eq 3 is

employed for both the neutral and the negatively charged system and an average or root-mean-square of the two criteria is minimized. The results for the tuned LC-PBE0 functional shown in Figure 5 suggest that it might be worthwhile to consider such a criterion in future studies of OR with range-separated hybrid functionals.

■ ASSOCIATED CONTENT

S Supporting Information. Tuning γ for norbornenone at the LC-BLYP and for the DFT-D3 structure of [7]helicene at the LC-PBE0 level. Tuning γ at the CAM-B3LYP level. The corresponding numerical values of $\Delta E(\gamma)$. Total energy of the C atom as a function of the electron number. Total energy of the β -pinene molecule as a function of the electron number. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: jochena@buffalo.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENT

This work has been supported by grant no. CHE 0952253 from the National Science Foundation. M.S. is grateful for financial support from the Foundation for Polish Science (“START” scholarship). The authors would like to acknowledge the Center for Computational Research (CCR) at the University at Buffalo for providing computational resources. The authors thank Dr. Niranjana Govind for technical advice and helpful discussions and Prof. Jeanne Crassous for information regarding the OR of 2-alkynyl[6]helicene.

■ REFERENCES

- (1) Parr, R. G.; Yang, W. *Density-Functional Theory of Atoms and Molecules*; Oxford University Press: New York, 1989.
- (2) Dreizler, R. M.; Gross, E. K. U. *Density Functional Theory. An Approach to the Quantum Many-Body Problem*; Springer-Verlag: New York, 1990.
- (3) Gross, E. K. U.; Dobson, J. F.; Petersilka, M. *Top. Curr. Chem.* **1996**, *181*, 81–172.
- (4) Engel, E.; Dreizler, R. M. *Density Functional Theory. An Advanced Course*; Springer-Verlag: Berlin, 2011.
- (5) Ziegler, T. *Chem. Rev.* **1991**, *91*, 651–667.
- (6) Seminario, J. M.; Politzer, P. *Modern Density Functional Theory. A Tool for Chemistry*; Elsevier: Amsterdam, 1995.
- (7) Seminario, J. M. *Recent Developments and Applications of Modern Density Functional Theory*; Elsevier: Amsterdam, 1996.
- (8) Siegbahn, P. E. M.; Blomberg, M. R. A. *Annu. Rev. Phys. Chem.* **1999**, *50*, 221–249.
- (9) van Doren, V.; van Alsenoy, C.; Geerlings, P. *Density Functional Theory and its Application to Materials: Antwerp, Belgium, June 8–10, 2000*; American Institute of Physics: Melville, NY, 2001.
- (10) Koch, W.; Holthausen, M. C. *A Chemist's Guide to Density-Functional Theory*; Wiley-VCH: New York, 2000.
- (11) Furche, F.; Ahlrichs, R. *J. Chem. Phys.* **2002**, *117*, 7433–7447.
- (12) Marques, M. A. L.; Ullrich, C. A.; Nogueira, F.; Rubio, A.; Burke, K.; Gross, E. K. U. *Time-Dependent Density Functional Theory*; Springer-Verlag: Berlin, 2006.

- (13) Autschbach, J. Spectroscopic Properties Obtained from Time-Dependent Density Functional Theory (TD-DFT). In *Encyclopedia of Inorganic Chemistry*; Wiley-VCH: New York, 2009. DOI: 10.1002/0470862106.ia600.
- (14) Cramer, C. J.; Truhlar, D. G. *Phys. Chem. Chem. Phys.* **2009**, *11*, 10757–10816.
- (15) Elliott, P.; Furche, F.; Burke, K. Excited States from Time-Dependent Density Functional Theory. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Cundari, T. R., Eds.; John Wiley & Sons, Inc.: Hoboken, NY, 2009; Vol. 26, DOI: 10.1002/9780470399545.ch3.
- (16) Autschbach, J.; Nitsch-Velasquez, L.; Rudolph, M. *Top. Curr. Chem.* **2011**, *298*, 1–98.
- (17) Cohen, A. J.; Mori-Sanchés, P.; Yang, W. *Science* **2008**, *321*, 792–794.
- (18) Zhang, Y.; Yang, W. *J. Chem. Phys.* **1998**, *109*, 2604–2608.
- (19) Champagne, B.; Perpète, E. A.; van Gisbergen, S. J. A.; Baerends, E.; Snijders, J. G.; Soubra-Ghaoui, C.; Robins, K. A.; Kirtman, B. *J. Chem. Phys.* **1998**, *109*, 10489–10498.
- (20) Dreuw, A.; Weisman, J. L.; Head-Gordon, M. *J. Chem. Phys.* **2003**, *119*, 2943–2946.
- (21) Tozer, D. J. *J. Chem. Phys.* **2003**, *119*, 12697–12699.
- (22) Peach, M. J. G.; Benfield, P.; Helgaker, T.; Tozer, D. J. *J. Chem. Phys.* **2008**, *128*, 044118.
- (23) Autschbach, J. *ChemPhysChem* **2009**, *10*, 1–5.
- (24) Perdew, J. P.; Zunger, A. *Phys. Rev. B* **1981**, *23*, 5048–5079.
- (25) van Leeuwen, R.; Baerends, E. J. *Phys. Rev. A* **1994**, *49*, 2421–2431.
- (26) Ullrich, C. A.; Gossmann, U. J.; Gross, E. K. U. *Phys. Rev. Lett.* **1995**, *74*, 872–875.
- (27) Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 10180–10189.
- (28) Iikura, H.; Tsuneda, T.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 3540–3544.
- (29) Tawada, Y.; Tsuneda, T.; Yanagisawa, S.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2004**, *120*, 8425–8433.
- (30) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51–57.
- (31) Baer, R.; Neuhauser, D. *Phys. Rev. Lett.* **2005**, *94*, 043002(1)–043002(4).
- (32) Gerber, I. C.; Ángyán, J. G. *Chem. Phys. Lett.* **2005**, *415*, 100–105.
- (33) Vydrov, O. A.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 234109(1)–234109(9).
- (34) Vydrov, O. A.; Heyd, J.; Krukau, A. V.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 074106(1)–074106(9).
- (35) 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*, 558–562.
- (36) Vydrov, O. A.; Scuseria, G. E.; Perdew, J. P. *J. Chem. Phys.* **2007**, *126*, 154109(1)–154109(9).
- (37) Livshits, E.; Baer, R. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2932–2941.
- (38) Chai, J.; Head-Gordon, M. *J. Chem. Phys.* **2008**, *128*, 084106(1)–084106(15).
- (39) Jiménez-Hoyos, C. A.; Janesko, B. G.; Scuseria, G. E. *J. Phys. Chem. A* **2009**, *113*, 11742–11749.
- (40) Steinmann, S. N.; Wodrich, M. D.; Corminboeuf, C. *Theor. Chem. Acc.* **2010**, *127*, 429–442.
- (41) Peach, M. J. G.; Cohen, A. J.; Tozer, D. J. *Phys. Chem. Chem. Phys.* **2006**, *8*, 4543–4549.
- (42) Lange, A. W.; Rohrdanz, M. A.; Herbert, J. M. *J. Phys. Chem. B* **2008**, *112*, 6304–6308.
- (43) Rohrdanz, M. A.; Herbert, J. M. *J. Chem. Phys.* **2008**, *129*, 034107(1)–034107(9).
- (44) Shcherbin, D.; Ruud, K. *Chem. Phys.* **2008**, *349*, 234–243.
- (45) Song, J.; Hirose, T.; Tsuneda, T.; Hirao, K. *J. Chem. Phys.* **2007**, *126*, 154105(1)–154105(7).
- (46) Stein, T.; Kronik, L.; Baer, R. *J. Am. Chem. Soc.* **2009**, *131*, 2818–2820.
- (47) Stein, T.; Eisenberg, H.; Kronik, L.; Baer, R. *Phys. Rev. Lett.* **2010**, *105*, 266802(1)–266802(4).
- (48) Kuritz, N.; Stein, T.; Baer, R.; Kronik, L. *J. Chem. Theory Comput.* **2011**, *7*, 2408–2415.
- (49) Baer, R.; Livshits, E.; Salzner, U. *Annu. Rev. Phys. Chem.* **2010**, *61*, 85–109.
- (50) Livshits, E.; Baer, R. *J. Phys. Chem. A* **2008**, *112*, 12789–12791.
- (51) Salzner, U.; Baer, R. *J. Chem. Phys.* **2009**, *131*, 231101(1)–231101(5).
- (52) Stein, T.; Kronik, L.; Baer, R. *J. Chem. Phys.* **2009**, *131*, 244119(1)–244119(5).
- (53) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. *J. Chem. Phys.* **2007**, *126*, 191109(1)–191109(5).
- (54) Perdew, J. P.; Parr, R. G.; Levy, M.; Balduz, J. L. *Phys. Rev. Lett.* **1982**, *49*, 1691–1694.
- (55) Tsuneda, T.; Song, J.; Suzuki, S.; Hirao, K. *J. Chem. Phys.* **2010**, *133*, 174101(1)–174101(9).
- (56) Polavarapu, P. L.; Chakraborty, D. K. *J. Am. Chem. Soc.* **1998**, *120*, 6160–6164.
- (57) Stephens, P. J.; Devlin, F. J.; Cheeseman, J. R.; Frisch, M. J.; Rosini, C. *Org. Lett.* **2002**, *4*, 4595–4598.
- (58) McCann, D. M.; Stephens, P. J.; Cheeseman, J. R. *J. Org. Chem.* **2004**, *69*, 8709–8717.
- (59) Giorgio, E.; Viglione, R. G.; Zanasi, R.; Rosini, C. *J. Am. Chem. Soc.* **2004**, *126*, 12968–12976.
- (60) Stephens, P. J.; McCann, D. M.; Cheeseman, J. R.; Frisch, M. J. *Chirality* **2005**, *17*, S52–S64.
- (61) McCann, D. M.; Stephens, P. J. *J. Org. Chem.* **2006**, *71*, 6074–6098.
- (62) Autschbach, J.; Jensen, L.; Schatz, G. C.; Tse, Y. C. E.; Krykunov, M. *J. Phys. Chem. A* **2006**, *110*, 2461–2473.
- (63) Stephens, P. J.; Devlin, F. J.; Cheeseman, J. R.; Frisch, M. J. *J. Phys. Chem. A* **2001**, *105*, 5356–5371.
- (64) Stephens, P. J.; Devlin, F. J.; Cheeseman, J. R.; Frisch, M. J.; Bortolini, O.; Besse, P. *Chirality* **2003**, *15*, S57–S64.
- (65) Crawford, T. D.; Stephens, P. J. *J. Phys. Chem. A* **2008**, *112*, 1339–1345.
- (66) Srebro, M.; Govind, N.; de Jong, W. A.; Autschbach, J. *J. Phys. Chem. A* **2011**, *115*, 10930–10949.
- (67) Richard, R. M.; Herbert, J. M. *J. Chem. Theory Comput.* **2011**, *7*, 1296–1306.
- (68) Bylaska, E. J.; de Jong, W. A.; Govind, N.; Kowalski, K.; Straatsma, T. P.; Valiev, M.; van Dam, H. J. J.; Wang, D.; Apra, E.; Windus, T. L.; Hammond, J.; Aquino, F.; Nichols, P.; Hirata, S.; Hackler, M. T.; Zhao, Y.; Fan, P.-D.; Harrison, R. J.; Dupuis, M.; Smith, D. M. A.; Glaesemann, K.; Nieplocha, J.; Tipparaju, V.; Krishnan, M.; Vazquez-Mayagoitia, A.; Jensen, L.; Swart, M.; Wu, Q.; Van Voorhis, T.; Auer, A. A.; Nooijen, M.; Crosby, L. D.; Brown, E.; Cisneros, G.; Fann, G. I.; Fruchtl, H.; Garza, J.; Hirao, K.; Kendall, R.; Nichols, J. A.; Tsemekhman, K.; Wolinski, K.; Anchell, J.; Bernholdt, D.; Borowski, P.; Clark, T.; Clerc, D.; Dachsel, H.; Deegan, M.; Dyall, K.; Elwood, D.; Glendening, E.; Gutowski, M.; Hess, A.; Jaffe, J.; Johnson, B.; Ju, J.; Kobayashi, R.; Kuttel, R.; Lin, Z.; Littlefield, R.; Long, X.; Meng, B.; Nakajima, T.; Niu, S.; Pollack, L.; Rosing, M.; Sandrone, G.; Stave, M.; Taylor, H.; Thomas, G.; van Lenthe, J.; Wong, A.; Zhang, Z. *NWChem, A Computational Chemistry Package for Parallel Computers*, version 6 (2011 developer's version); Pacific Northwest National Laboratory: Richland, WA, 2011.
- (69) Valiev, M.; Bylaska, E. J.; Govind, N.; Kowalski, K.; Straatsma, T. P.; van Dam, H. J. J.; Wang, D.; Nieplocha, J.; Apra, E.; Windus, T. L.; de Jong, W. A. *Comput. Phys. Commun.* **2010**, *181*, 1477–1489.
- (70) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007–1023.
- (71) Kendall, R. A.; Dunning, T. H.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796–6806.
- (72) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

- (73) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1997**, *78*, 1396.
- (74) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.
- (75) Ernzerhof, M.; Scuseria, G. E. *J. Chem. Phys.* **1999**, *110*, 5029–5036.
- (76) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.
- (77) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.
- (78) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200–1211.
- (79) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623–11627.
- (80) Autschbach, J. *Comput. Lett.* **2007**, *3*, 131–150.
- (81) Autschbach, J. *ChemPhysChem* **2011**, *12*, 3224–3235.
- (82) Müller, T.; Wiberg, K. B.; Vaccaro, P. H. *J. Phys. Chem. A* **2000**, *104*, 5959–5968.
- (83) Müller, T.; Wiberg, K. B.; Vaccaro, P. H.; Cheeseman, J. R.; Frisch, M. J. *J. Opt. Soc. Am. B* **2002**, *19*, 125–141.
- (84) Wilson, S. M.; Wiberg, K. B.; Cheeseman, J. R.; Frisch, M. J.; Vaccaro, P. H. *J. Phys. Chem. A* **2005**, *109*, 11752–11764.
- (85) Mach, T. J.; Crawford, T. D. *J. Phys. Chem. A* **2011**, *115*, 10045–10051.
- (86) Pedersen, T. B.; Koch, H.; Boman, L.; de Merás, A. M. *J. S. Chem. Phys. Lett.* **2004**, *393*, 319–326.
- (87) Tam, M. C.; Russ, N. J.; Crawford, T. D. *J. Chem. Phys.* **2004**, *121*, 3550–3557.
- (88) Kongsted, J.; Pedersen, T. B.; Strange, M.; Osted, A.; Hansen, A. E.; Mikkelsen, K. V.; Pawłowski, F.; Jørgensen, P.; Hättig, C. *Chem. Phys. Lett.* **2005**, *401*, 385–392.
- (89) Kumata, Y.; Furukawa, J.; Fueno, T. *Bull. Chem. Soc. Jpn.* **1970**, *43*, 3920–3921.
- (90) Lightner, D. A.; Gawronski, J. K.; Bouman, T. D. *J. Am. Chem. Soc.* **1980**, *102*, 5749–5754.
- (91) Brown, H. C.; Zaidlewicz, M.; Bhat, K. S. *J. Org. Chem.* **1989**, *54*, 1764–1766.
- (92) Ruud, K.; Zanasi, R. *Angew. Chem., Int. Ed.* **2005**, *44*, 3594–3596.
- (93) Crawford, T. D.; Tam, M. C.; Abrams, M. L. *Mol. Phys.* **2007**, *105*, 2607–2617.
- (94) Mort, B. C.; Autschbach, J. *Chem. Phys. Chem* **2008**, *9*, 159–170.
- (95) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, Revision A.1; Gaussian, Inc.: Wallingford, CT, 2009.
- (96) Woon, D. E.; Dunning, T. H. *J. Chem. Phys.* **1994**, *100*, 2975–2988.
- (97) Moscovitz, A. *Adv. Chem. Phys.* **1962**, *4*, 67–112.
- (98) Wiberg, K. B.; Wang, Y. G.; Wilson, S. M.; Vaccaro, P. H.; Cheeseman, J. R. *J. Phys. Chem. A* **2006**, *110*, 13995–14002.
- (99) Ruud, K.; Stephens, P. J.; Devlin, F. J.; Taylor, P. R.; Cheeseman, J. R.; Frisch, M. J. *Chem. Phys. Lett.* **2003**, *373*, 606–614.
- (100) Autschbach, J.; Patchkovskii, S.; Ziegler, T.; van Gisbergen, S. J. A.; Baerends, E. J. *J. Chem. Phys.* **2002**, *117*, 581–592.
- (101) Krykunov, M.; Autschbach, J. *J. Chem. Phys.* **2005**, *123*, 114103–10.
- (102) Mort, B. C.; Autschbach, J. *J. Phys. Chem. A* **2005**, *109*, 8617–8623.
- (103) Christiansen, O.; Koch, H.; Jørgensen, P. *Chem. Phys. Lett.* **1995**, *243*, 409–418.
- (104) Wong, B. M.; Hsieh, T. H. *J. Chem. Theory Comput.* **2010**, *6*, 3704–3712.
- (105) Murphy, V. L.; Kahr, B. *J. Am. Chem. Soc.* **2011**, *133*, 12918–12921.
- (106) Lopata, K.; Reslan, R.; Kowalska, M.; Neuhauser, D.; Govind, N.; Kowalski, K. *J. Chem. Theory Comput.* **2011**, *7*, 3686–3693.
- (107) Anger, E.; Srebro, M.; Vanthuyne, N.; Toupet, L.; Roussel, C.; Autschbach, J.; Crassous, J.; Réau, R. 2011, in preparation.
- (108) Martin, R. H.; Flammang-Barbieux, M.; Cosyn, J. P.; Gelbcke, M. *Tetrahedron Lett.* **1968**, *9*, 3507–3510.
- (109) Newman, M. S.; Lutz, W. B.; Lednicer, D. *J. Am. Chem. Soc.* **1955**, *77*, 3420–3421.
- (110) Lightner, D. A.; Hefelfinger, D. T.; Powers, T. W.; Frank, G. W.; Trueblood, K. N. *J. Am. Chem. Soc.* **1972**, *94*, 3492–3497.
- (111) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *J. Chem. Phys.* **2010**, *132*, 154104–154119.
- (112) Grimme, S.; Mück-Lichtenfeld, C.; Antony, J. *J. Phys. Chem. C* **2007**, *111*, 11199–11207.
- (113) Norel, L.; Rudolph, M.; Vanthuyne, N.; Williams, J. A. G.; Lescop, C.; Roussel, C.; Autschbach, J.; Crassous, J.; Réau, R. *Angew. Chem., Int. Ed.* **2010**, *49*, 99–102.
- (114) Anger, E.; Rudolph, M.; Norel, L.; Zrig, S.; Shen, C.; Vanthuyne, N.; Toupet, L.; Williams, J. A. G.; Roussel, C.; Autschbach, J.; Crassous, J.; Réau, R. *Chem.—Eur. J.* **2011**, *17*, 14178–14198.
- (115) Perdew, J. P.; Levy, M. *Phys. Rev. Lett.* **1983**, *51*, 1884–1887.
- (116) Mori-Sánchez, P.; Cohen, A. J.; Yang, W. *Phys. Rev. Lett.* **2008**, *100*, 146401(1)–146401(4).

Encapsulation Influence on EPR Parameters of Spin-Labels: 2,2,6,6-Tetramethyl-4-methoxypiperidine-1-oxyl in Cucurbit[8]uril

Zilvinas Rinkevicius,^{*,†,§} Bogdan Frecuş,[†] N. Arul Murugan,[†] Olav Vahtras,[†] Jacob Kongsted,[‡] and Hans Ågren[†]

[†]Department of Theoretical Chemistry & Biology, School of Biotechnology, Royal Institute of Technology, SE-106 91 Stockholm, Sweden

[‡]Department of Physics, Chemistry and Pharmacy, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark

[§]Swedish e-Science Research Center (SeRC), Royal Institute of Technology, SE-100 44 Stockholm, Sweden

ABSTRACT: Encapsulation of a nitroxide spin label into a host cavity can prolong the lifetime of the spin label in biological tissues and other environments. Although such paramagnetic supramolecular complexes have been extensively studied experimentally, there is yet little understanding of the role of the encapsulation on the magnetic properties of the spin labels and their performance at the atomistic level. In this work, we approach this problem by modeling encapsulation induced changes of the magnetic properties of spin labels for a prototypical paramagnetic guest–host complex, 2,2,6,6-tetramethyl-4-methoxypiperidine-1-oxyl, enclosed in the hydrophobic cavity of cucurbit[8]uril, using state-of-the-art hybrid quantum mechanics/molecular mechanics methodology. The results allow a decomposition of the encapsulation shift of the electronic *g*-tensor and the nitrogen isotropic hyperfine coupling constant of nitroxide radical into a set of distinct contributions associated with the host cavity confinement and with changes of the local solvent environment of the spin label upon encapsulation. It is found that the hydrophobic cavity of cucurbit[8]uril only weakly influences the electronic *g*-tensor of the 2,2,6,6-tetramethyl-4-methoxypiperidine-1-oxyl but induces a significant encapsulation shift of the nitrogen hyperfine coupling constant. The latter is caused by the change of topology of the hydrogen bonding network and the nature of the hydrogen bonds around the spin label induced by the hydrophobic cavity of the inclusion host. This indirect effect is found to be more important than the direct influence of the cavity exerted on the radical. The ramification of this finding for the use of approximate methods for computing electron paramagnetic resonance spectra of spin labels and for designing optimal spin labels based on guest–host templates is discussed.

1. INTRODUCTION

Electron paramagnetic resonance (EPR) spectroscopy is a most versatile technique for studies of biomolecules in controlled environments,^{1–12} like solvents or crystals. This technique has extensively been used for exploitation of structure; surface properties; and dynamics of proteins, membranes, and other biological complexes, prevalently employing nitroxides as spin labels owing to their chemical stability.^{1–8,11,12} However, only a handful of such studies have yet been carried out in native biological environments,^{13–18} *in vivo* or *in vitro*, due to the reduced stability of nitroxides or other spin labels in such environments.

The major chemical obstacles for *in vivo* EPR spin labeling studies are the rapid one-electron reduction of the nitroxide radical to the corresponding EPR silent hydroxylamine^{19–22} as well as the two-electron cellular bioreduction which may occur following the oxidation of the stable spin label.²³ Moreover, the reduction of nitroxides in biological tissues depends on the concentration of oxygen and endogenous reducing agents, such as glutathione,^{19,24} and on the spin label structure (piperidine vs pyrrolidine ring).²⁵

A way to overcome these problems is to increase the steric hindrance around the spin label, thus employing a molecular complex which encapsulates the nitroxide into a protective cavity that can effectively increase the time during which the EPR signal can be detected.^{20–22} However, the inclusion of the nitroxide spin label into a protective host cavity introduces the additional

complexity that the spectral changes of the EPR signal caused by the host cavity, the solvent, and the target system, cannot be disentangled without microscopic knowledge of the interactions involved. This work is an attempt to resolve this matter and to advance the understanding of the mechanisms governing the changes of the EPR spin Hamiltonian parameters of spin labels under encapsulation. For this purpose, we apply state-of-the-art molecular modeling techniques comprising hybrid quantum mechanics/molecular mechanics for the evaluation of EPR spin Hamiltonian parameters,^{26–29} and we study a prototypical paramagnetic guest–host system, 2,2,6,6-tetramethyl-4-methoxypiperidine-1-oxyl in cucurbit[8]uril, which has been recently extensively studied using experimental EPR methods.²⁰ This is part of an effort to design more reliable procedures for EPR spectra analysis, which takes into account various microscopic mechanisms responsible for spin Hamiltonian parameters, being applicable, in addition to “guest–host” complexes, to spin-labels restrained within hydrophobic cavities on protein surfaces.

2. COMPUTATIONAL DETAILS

In this work, an integrated approach^{30,31} has been employed to evaluate EPR spin Hamiltonian parameters of the 2,2,6,6-tetramethyl-4-methoxypiperidine-1-oxyl (4M) radical in either

Received: November 14, 2011

Published: December 07, 2011

aqueous solution or encapsulated within the molecular cavity of cucurbit[8]uril (CB[8]) also embedded in an aqueous solution. This approach consists of two steps—classical molecular dynamics (MD) simulations of the solute in its solvent environment and subsequent hybrid quantum mechanics/molecular mechanics (QM/MM) methodology based calculations of EPR spin Hamiltonian parameters over the set of uncorrelated snapshots extracted from the MD trajectory. In the following, we will describe the technical details of both computational steps of this integrated approach, which has been used in this work to study encapsulation effects on the EPR spin Hamiltonian parameters of the 4M radical.

2.1. Classical Molecular Dynamics Simulations. In this work we carried out two separate molecular dynamics simulations at ambient temperature: one for the 4M radical in aqueous solution and a second for the 4M encapsulated in the cavity of CB[8], a guest–host complex in aqueous solution. In the first MD simulation, the 4M radical was solvated in an orthorhombic box with dimensions of approximately 69.6, 68.1, and 65.4 Å and which contained 10 192 water solvent molecules. In the second MD simulation, the solvent box dimension was approximately 76.4, 71.5, and 75.8 Å and contained 13 567 water molecules as well as the 4M@CB guest–host complex. Both simulations have been performed within the isothermal–isobaric ensemble, and the temperature and pressure have been controlled by connecting the simulation box to a thermostat and a barostat.^{32–34} The MD simulations have been carried out using the AMBER molecular dynamics package.³⁵ Concerning the force fields used in the MD simulations, we used the TIP3P³⁶ force field to describe the water molecules and the GAFF³⁷ force field for the CB[8] molecule. The CB[8] atomic charges have been derived using the CHELPG³⁸ procedure at the B3LYP/6-311++G(d,p) level. In addition to this rather conventional force field choice, we have faced the challenge of selecting a suitable force field for the 4M radical, which would provide a reliable description of the structural parameters of the R₂NO[•] moiety in the 4M radical. After evaluating several alternative force fields designed for description of the nitroxides and spin labels, we settled on a recently developed extension of the AMBER force field by Barone et al.³⁹ A time step of 1 fs has been chosen for the integration of the equation of motion. Using the above outlined setup, the MD simulations have been carried out for the free 4M radical and the 4M@CB[8] guest–host complex in aqueous solution, where the production trajectory length was set to 0.5 ns after an equilibration run of 0.5 ns. From both MD simulations, we extracted 100 snapshots, which have been taken with regular 5 ps time intervals from the production MD trajectory, for subsequent QM/MM calculations of EPR spin Hamiltonian parameters in the second step of the integrated approach for molecular properties modeling.

2.2. Hybrid QM/MM Calculations of EPR Spin-Hamiltonian Parameters. The EPR spin-Hamiltonian parameters of the 4M radical, i.e., the electronic g-tensor and nitrogen isotropic hyperfine coupling constant (hfcc), have in this work been computed using the hybrid density functional theory/molecular mechanics approach²⁸ and the hybrid density functional restricted–unrestricted theory/molecular mechanics approach,²⁹ respectively. In these calculations, we employed a well established setup for evaluation of both parameters, which have been extensively benchmarked in our previous works on the prototypical system, di-tert-butyl-nitroxide in aqueous solution.^{28,29} Thus, according to the methodology suggested in our previous works,^{28,29} we

carried out electronic g-tensor and nitrogen isotropic hfcc calculations limiting the QM region to the 4M radical and treating the CB[8] molecule and all molecules within a 20 Å radius around the 4M radical as the MM region. The QM region has been described at the B3LYP^{40–43} level using the Huz-III basis set^{44,45} for electronic g-tensor calculations, while in the case of the nitrogen isotropic hfcc calculations we have used the more flexible core region basis set Huz-III_{su}.^{44,45} Here, we would like to point out that the selection of the B3LYP functional for our calculations is motivated by our desire to have the same level description of the 4M radical during evaluation of both spin-Hamiltonian parameters, as this allows us to compare the encapsulation effect on both spin Hamiltonian parameters more fairly. However, as we already noted in our previous works,^{28,29} a more accurate description of the electronic g-tensor and nitrogen isotropic hfcc of the 4M radical can be obtained by using the BP86^{41,46} or PBE0^{47–50} functionals, respectively. After outlining the technical details of the QM region description of the hybrid QM/MM calculations, let us turn to the second important technical aspect of these calculations, namely, the description of the MM region. For the water molecules in the MM region, we used the MM-3 force field,²⁸ which has shown very good performance in our previous works^{28,29} and is thus expected to perform well in representing the aqueous environment of both the 4M radical and the 4M@CB[8] guest–host complex. The single remaining molecule in our MM region, namely, the CB[8] host molecule, has in all calculations been described using the same level of force field as the water molecules; i.e., the force field parametrization included point charges, distributed dipoles and quadrupoles, and distributed anisotropic polarizabilities. This force field has been generated following the LoProp procedure⁵¹ at the B3LYP/6-31+G(d) level of theory and has been evaluated for each snapshot separately, thereby including molecular distortions in the CB[8] guest. All of the above outlined calculations have been carried out using the development version of the DALTON quantum chemistry program.⁵²

3. RESULTS AND DISCUSSION

EPR spin Hamiltonian parameters of nitroxide spin labels show a remarkable dependence on the geometrical structure of the R₂NO[•] moiety and its immediate environment. These features have, over the years, been the subject of numerous experimental and theoretical studies,^{28,29,31,53–72} and by and large, the behavior of nitroxides in protic and aprotic solvents is now quite well understood. However, guest–host complexes, in which nitroxides are included in various hollow compounds like cucurbiturils, have been less extensively investigated,^{20,21,73–77} and so far no studies have addressed the changes of EPR spin Hamiltonian parameters upon encapsulation of nitroxide in host cavities at the microscopic level. In order to resolve this matter and to investigate the atomistic origin of these effects, we study a prototype guest–host system (see Figure 1) consisting of the 2,2,6,6-tetramethyl-4-methoxypiperidin-1-oxyl (4M) radical and cucurbit[8]uril (CB[8]) solvated in aqueous solution. In the following, we describe the effect of encapsulation in CB[8] on the 4M internal structural parameters and the local solvation environment of the R₂NO[•] moiety. Using the results for the structural parameters of the solvated 4M@CB[8] complex, it is possible to dismantle the encapsulation induced shift of the electronic g-tensor and the isotropic hyperfine coupling constant (hfcc) of nitrogen into distinct contributions due to the CB[8] cavity and

the bulk water solvent and into the more indirect contributions from the interplay between solvation and encapsulation effects.

The spin Hamiltonian parameters, namely, the electronic g -tensor and the nitrogen isotropic hyperfine coupling constant, of nitroxides are mainly governed by the structure of two orbitals:^{28,29,53,59,62} the doubly occupied n -type HOMO, in which the oxygen lone pair resides, and the singly occupied π -type SOMO, which holds the unpaired electron (see Figure 1). These two orbitals are localized on the R_2NO^\bullet moiety, and the geometrical structure as well as the local solvent environment of this moiety are therefore the key factors determining the values of the spin Hamiltonian parameters. Taking this into account, the first step toward understanding the influence of encapsulation on the guest spin Hamiltonian parameters is to examine the

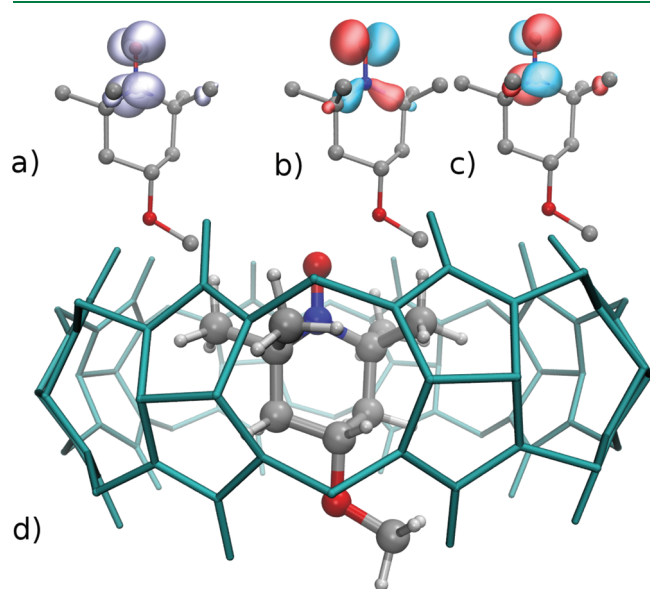


Figure 1. Graphical illustration of (a) spin density, (b) HOMO, and (c) SOMO of the 4M radical, and its inclusion complex with the CB[8] host (hydrogen atoms have been removed from CB[8] for clarity).

differences of the internal 4M radical structural parameters—the NO bond distance and improper dihedral angle θ (see Figure 2 for the definition of this angle)—between the two systems: the free 4M radical and the 4M@CB[8] complex solvated in water. The results of molecular dynamic simulations indicate that the dynamics of the NO bond in terms of both bond length and out-of-plane movement are almost the same for the two systems, see Figure 2. Because of the limited encapsulation effect on the dynamics of the internal 4M radical, we can expect the changes of the spin Hamiltonian parameters induced by the host cavity to be negligible. However, differently from the internal geometrical structure of the 4M radical, the encapsulation in CB[8] has a profound effect on the topology and dynamics of the local solvation of the R_2NO^\bullet moiety in the 4M radical, where the averaged number of hydrogen bonded water molecules to the R_2NO^\bullet oxygen decreases from 1.8 to 1.3 going from the free 4M nitroxide to the 4M@CB[8] complex in aqueous solution. Furthermore, in addition to the conventional hydrogen bonding topology “ R_2NO^\bullet moiety \cdots water \cdots water” two new topologies are present in the aqueous 4M@CB[8] complex (see Figure 3): “ R_2NO^\bullet moiety \cdots water \cdots CB[8]”, where one water hydrogen bonds to both 4M and CB[8], and “ R_2NO^\bullet moiety \cdots water \cdots water \cdots CB[8]”, where two waters form a hydrogen bond bridge between 4M and CB[8]. As one can see from panels c and d in Figure 3, the encapsulation of the 4M radical in CB[8] not only reduces the average number of water molecules bound to it but also changes the distribution of these waters over the MD trajectory; fewer configurations without hydrogen bonded waters to the 4M radical appear and the number of configurations with three hydrogen bonded waters to the 4M radical decreases significantly. The hydrogen bonds formed in the two cases also show some peculiarities. In the case of the 4M radical in aqueous solution, all waters hydrogen bonded to the 4M radical follow the conventional “ R_2NO^\bullet moiety \cdots water \cdots water” topology. Upon encapsulation of the 4M radical in CB[8], only around 4% of the hydrogen bonded water molecules retain the same hydrogen bonding topology as in case of the free 4M radical in aqueous solution, while 96% of the hydrogen bonded waters adopt the above-mentioned

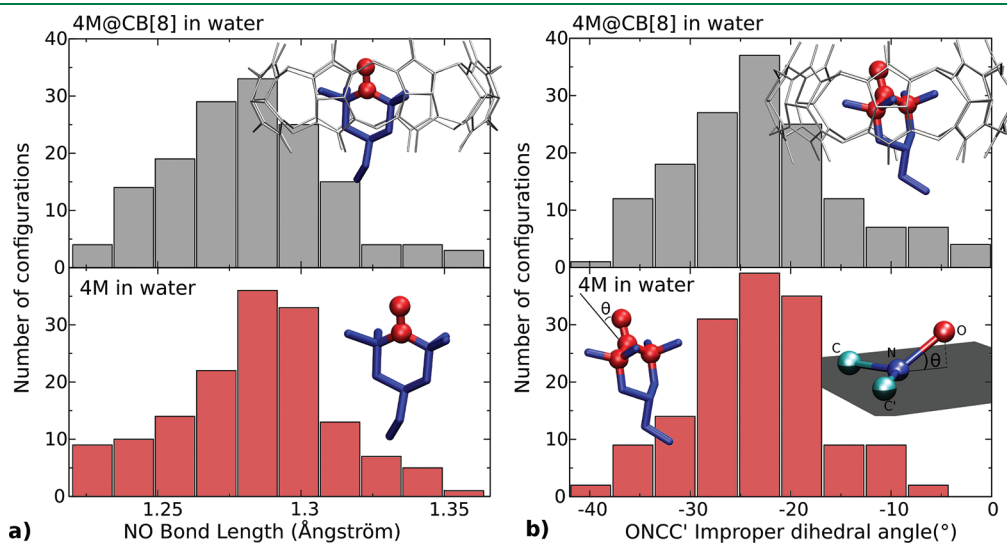


Figure 2. Internal parameters of the 4M radical and 4M@CB[8] complex in aqueous solution: (a) NO distance histograms for both MD simulations: top, 4M@CB[8] in water; bottom, -4 M in water. (b) NOCC' improper dihedral angle histograms for both MD simulations: top, -4 M@CB[8] in water; bottom, 4 M in water.

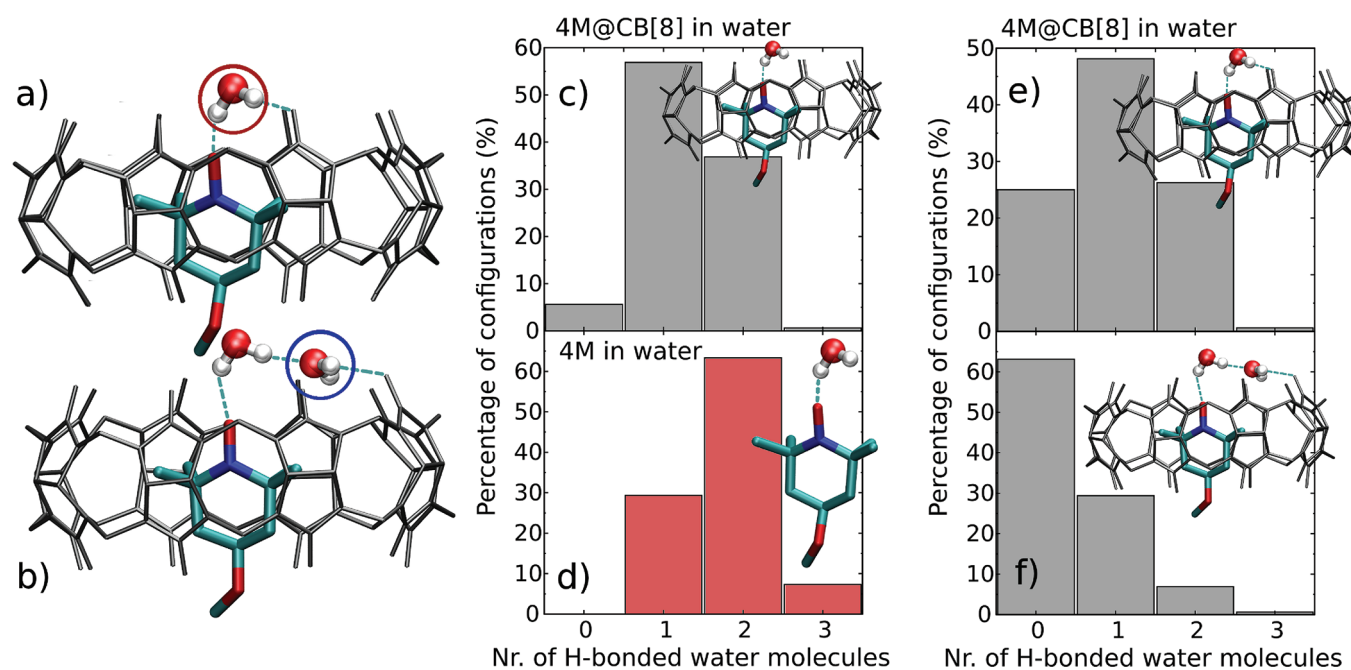


Figure 3. Graphical illustration of the supramolecular assemblies of 4M@CB[8] (hydrogen atoms have been removed for clarity) highlighting the two possible topologies that involve hydrogen-bonding solvent molecules to the guest and host systems: (a) “R₂NO• moiety···water···CB[8]” and (b) “R₂NO• moiety···water···water···CB[8]”. Hydrogen bond distribution (c) 4M@CB[8] in water, (d) 4M in water MD simulations. Distribution of new hydrogen bonding topologies in the 4M@CB[8] complex in aqueous solution: (e) “R₂NO• moiety···water···CB[8]” and (f) “R₂NO• moiety···water···water···CB[8]”.

Table 1. Decomposition of the Shift in the Electronic g-Tensor and Nitrogen Isotropic hfcc in Aqueous 4M Radical Encapsulated in CB[8] Host Cavity Based on the Hybrid QM/MM Calculations^a

origin of contribution	Δg_{iso} in ppm	a_N in Gauss
internal dynamics of R ₂ NO• moiety ^b	−61	−0.01
interaction with water molecules ^c	22	−0.51
interaction with CB[8] ^d	12	−0.05
changes in hydrogen bond strength due to solvent and CB[8] interaction ^e	−37	0.12
total ^f	−64	−0.45
exptl. ^g	−240	−0.41

^a EPR spin-Hamiltonian parameters computed at the Huz-III/B3LYP (g-tensor) and HUZ-IIIu3/B3LYP (hfcc) levels of theory. ^b Determined as the difference between averaged vacuum values of the EPR spin-Hamiltonian parameter, which have been computed over 100 snapshots with geometries for each snapshot taken from the MD simulations of the aqueous 4M@CB[8] complex and of the aqueous radical. ^c Determined as the difference between averaged values of EPR spin-Hamiltonian parameter in aqueous solution with internal dynamics of the R₂NO• moiety subtracted. ^d Determined as the difference between averaged values of the EPR spin-Hamiltonian parameter computed for free and encapsulated 4M radical in a vacuum with snapshot geometries taken from the 4M@CB[8] complex in aqueous solution MD trajectory. ^e Determined as the difference between averaged values of the EPR spin-Hamiltonian parameter computed for the 4M radical in aqueous solution with and without a CB[8] host included in QM/MM calculations. ^f Sum of all contributions to the encapsulation shift of the EPR spin-Hamiltonian parameter. ^g Experimental EPR parameters of 4M radical in the presence of 2 mM sodium ascorbate and 3.4 mM CB[8] with the pH adjusted with LiOH, obtained by Bardelang et al.,⁴ more specifically, the difference between the values reported for the aqueous solutions of the encapsulated 4M radical and free 4M radical.

two new topologies involving the CB[8] host cavity. We can expect this significant alternation of local solvation of the R₂NO• moiety upon encapsulation to be the main mechanism responsible for changes of its spin Hamiltonian parameters.

In order to shed light on the mechanism for the encapsulation induced shift of the electronic g-tensor and the nitrogen isotropic hfcc, we decompose the encapsulation shift into contributions of different physical origin. According to the MD simulations of the free 4M radical and the solvated 4M@CB[8] complex, we identified the following possible mechanisms for the encapsulation shift: (a) alternation of

internal dynamics of the R₂NO• moiety in the 4M radical, (b) reduction of the average number of water molecules bonded to the oxygen of R₂NO•, and (c) alternation of the hydrogen bond strength between R₂NO• and water molecules due to changes in the hydrogen bonding topology. In addition to these structural mechanisms, we also consider the direct effect of the CB[8] cavity on the electronic structure of the 4M radical and its EPR spin Hamiltonian parameters. The hybrid QM/MM computational results of this decomposition of the encapsulation shift of Δg_{iso} and a_N are tabulated in Table 1.

Overall, the QM/MM results qualitatively reproduce the observed encapsulation shifts of the spin Hamiltonian parameters of the 4M radical. For the electronic g -tensor, our calculations underestimate the experimentally observed decrease of Δg_{iso} upon the 4M radical encapsulation in CB[8]. The reason can most likely be traced to differences in conditions for the experiment and our MD simulations. In fact, by their design, the MD simulations eliminate the influence of various external factors on the 4M radical aqueous environment, which can influence experimental results and provide a more clear foundation for evaluating the encapsulation effect on the spin Hamiltonian parameters. We therefore expect our QM/MM modeling results to be representative of the influence exerted of the CB[8] cavity on the 4M radical electronic structure and its magnetic properties in aqueous solution.

The four above identified mechanisms for the encapsulation shift of Δg_{iso} are evidently of different importance. According to the results in Table 1, the change of the internal 4M radical dynamics upon encapsulation in the CB[8] cavity is responsible for the decrease of Δg_{iso} by 61 ppm and is apparently the largest contribution to the encapsulation shift of Δg_{iso} . This is a rather unexpected result, since the MD simulations indicate that the internal parameters of the $\text{R}_2\text{NO}^\bullet$ moiety as well as its dynamics are only minorly altered upon encapsulation in the CB[8] cavity. Out of the remaining mechanisms, the smallest contribution arises from the direct influence of the CB[8] cavity on the electronic structure of the 4M radical, which is approximately 5 times smaller than the previously discussed contribution. Furthermore, it acts in the opposite way, i.e. induces an increase of Δg_{iso} upon encapsulation (see Table 1 for details). The last two mechanisms responsible for the encapsulation shift of Δg_{iso} are related to changes of the aqueous environment of the 4M radical going from its free to its guest–host complex form. The “interaction with water molecules” mechanism includes contributions from alternation of the aqueous environment structure (waters hydrogen bonded to the 4M radical as well as waters in the bulk of the aqueous solution) upon encapsulation, which increases Δg_{iso} , as can be seen from Table 1. This is caused by the number of hydrogen bonded waters to the $\text{R}_2\text{NO}^\bullet$ moiety in the aqueous 4M@CB[8] complex, which in turn leads to a smaller blue shift of the $n \rightarrow \pi$ excitation (see Figure 1) and consequently to a larger Δg_{iso} of the encapsulated 4M radical compared to its free form in aqueous solution. The final mechanism of the four is the change of the hydrogen bond strength between oxygen of the $\text{R}_2\text{NO}^\bullet$ moiety and the water molecules, induced by the interaction of the water molecules with the CB[8] cavity. As we already established, most of the hydrogen bonds between the 4M radical and the waters in the 4M@CB[8] complex undergo an alternation of their topology, and we can thus expect this mechanism to give a significant contribution to the encapsulation shift of Δg_{iso} . The QM/MM results in Table 1 verify this assumption and show that the alternation of hydrogen bond strengths due to the water interaction with CB[8] leads to almost a doubling of the encapsulation shift of Δg_{iso} compared to the one induced by the reduction of the averaged number of hydrogen bonded waters to the 4M radical upon encapsulation. Furthermore, these two encapsulation shifts of Δg_{iso} have opposite sign, and thus their overall relative importance for the total encapsulation shift of Δg_{iso} is diminished.

Taking into account the size of the four individual contributions to the encapsulation shift of Δg_{iso} , we can establish their decreasing order of importance: “alternation of the internal

Table 2. Electronic g -Tensor and Nitrogen Isotropic Hyperfine Coupling Constants of Free 4M Radical in Water and in 4M Radical in 4M@CB[8] Complex in Water^a

	Δg_{iso} in ppm	a_N in G
4M (vacuum) ^b	4063	14.85
4M (water) ^b	3705	17.00
4M (water) Exp. ^c	4201	17.04
4M (vacuum) ^d	4003	14.84
4M@CB[8] (water) ^d	3641	16.55
4M@CB[8] (water) exptl. ^c	3961	16.63

^aEPR spin-Hamiltonian parameters computed at the Huz-III/B3LYP (g-tensor) and HUZ-IIIsu3/B3LYP (hfcc) levels of theory. ^bStructural data taken from aqueous 4M radical MD simulation. ^cExperimental data taken from the Supporting Information of ref 20. ^dStructural data taken from the aqueous 4M radical encapsulated in CB[8] MD simulation.

dynamics of the 4M radical upon encapsulation” > “reduction of the averaged number of hydrogen bonded waters to the $\text{R}_2\text{NO}^\bullet$ moiety open encapsulation” + “change of the hydrogen bond nature and topology upon encapsulation” > “direct influence of the CB[8] cavity on the electronic structure of the 4M radical upon encapsulation”. We conclude that the encapsulation induced shift of the g -tensor is rather small (−64 ppm, see Table 1) and is significantly smaller than the shift induced by solvation of the 4M radical in water as shown by the data in Table 2. Therefore, the encapsulation of the 4M radical in the CB[8] cavity in aqueous solution only slightly affects the electronic g -tensor of the radical, indicating that it can be disregarded in practical analysis of EPR spectra of guest–host systems of such a kind and probably also for spin-labels residing in hydrophobic cavities of proteins. Thus, the QM/MM modeling of the 4M radical g -tensors provides for the first time theoretical support to the common assumption in empirical models that the electronic g -tensor of the spin label remains unchanged upon encapsulation.

For the second important spin Hamiltonian parameter—the nitrogen isotropic hyperfine coupling constant a_N —we find, in agreement with previous studies of nitroxides, that the local solvent environment effect on the 4M radical is larger than that for the electronic g -tensor. The QM/MM results tabulated in Table 1 indicate that the encapsulation of the 4M radical in the CB[8] cavity reduces the nitrogen isotropic hfcc by the significant amount of 0.45 G. Comparing theoretical results and experimental data, the former overestimate the encapsulation shift of a_N , probably for the same reasons as in the case of the electronic g -tensor, i.e., a mismatch between environmental conditions for the MD simulations and the ones encountered in the real experiments. Following a similar procedure as for Δg_{iso} , we decomposed the encapsulation shift of a_N into four contributions. The first most striking difference between the electronic g -tensor and the nitrogen isotropic hfcc is the small contribution to the latter from the changes in internal structure and dynamics of the 4M radical, while the opposite is found in the former case. This is in line with our expectations based on the analysis of MD results, which show that the $\text{R}_2\text{NO}^\bullet$ moiety changes only slightly going from free 4M radical to its encapsulated form in the aqueous environment. The reason can be traced to the electron spin density, which defines the value of a_N and which is solely localized on the $\text{R}_2\text{NO}^\bullet$ moiety (see Figure 1). The negligible geometry change in this moiety naturally translates into

a small encapsulation shift induced via this mechanism. Concerning the direct influence of the CB[8] cavity on the 4M electronic structure, our results for the two spin Hamiltonian parameters agree well, indicating a negligible size of this contribution in both cases. The two remaining contributions to the encapsulation shift of the nitrogen isotropic hfcc are associated with changes of the aqueous environment of the 4M radical and play, as we expect, a most important role. Among these two contributions, the large change of a_N is caused by the decrease of the averaged number of water molecules bonded to the R_2NO^\bullet moiety, while the effect of new hydrogen bonding topologies on the encapsulation shift is less pronounced, being almost 5 times smaller. Taking these results into account, we can establish the following order of the contributions to the encapsulation shift of a_N in decreasing importance: “reduction of the averaged number of hydrogen bonded waters to the R_2NO^\bullet moiety upon encapsulation” > “change of the hydrogen bond topology upon encapsulation” > “direct influence of the CB[8] cavity on the electronic structure of the 4M radical upon encapsulation” > “alternation of the internal dynamics of the 4M radical upon encapsulation”. This order of the contributions is in good agreement with the one predicted on the basis of our analysis of the MD simulations and indicates that the rather unexpected behavior of the encapsulation shift of the electronic g -tensor should be qualified by the overall small size of the contributions making it up. In total, according to results presented in Table 2, the solvent shift of a_N is 2.14 G going from a vacuum to aqueous solution, and the encapsulation shift, being 0.45 G, thus constitutes almost 25% of the solvent shift and cannot be neglected in the analysis of experimental EPR spectra. We conclude that encapsulation produces a significant effect on this spin Hamiltonian parameter, something that can be expected to be encountered in other similar “guest–host” complexes as well as in the case of spin-labels residing in hydrophobic cavities of proteins.

4. CONCLUSIONS

In the present work, we have given a theoretical perspective on the effect of encapsulation on the magnetic properties of spin labels encased in hydrophobic host cavities consisting of a protective shell molecule. We approached this problem by employing state-of-the-art hybrid quantum mechanics/molecular mechanics, which allow for a consistent description of spin labels in different environments. We studied the encapsulation effect of the spin label electronic g -tensor and the nitrogen isotropic hyperfine coupling constant in a prototypical guest–host system, consisting of 2,2,6,6-tetramethyl-4-methoxypiperidine-1-oxyl and cucurbit[8]uril, in aqueous solution. From our modeling results, several conclusions could be drawn on the physical origin of the EPR parameters of the spin labels and their dependence on encapsulation and solvation. It is shown that the hydrophobic cavity of CB[8] and other similar hosts, like cyclodextrines, only weakly influences the electronic g -tensor of the nitroxide but induces a noticeable encapsulation shift of the nitrogen hyperfine coupling constant. This finding provides for the first time theoretical support for the common assumption used in most empirical models that the electronic g -tensor of spin-labels remains unchanged upon encapsulation. However, the same assumption does not hold for the nitrogen isotropic hyperfine coupling constants, which experience significant encapsulation shifts.

The main difference between the spectroscopic properties obtained for the nitroxides in guest–host complexes and for the nitroxides in solution is attributed to the steric hindrance of the nitroxides into the hydrophobic cavity, which affects the local hydrogen bonding of the solvent molecules. Thus, upon encapsulation, fewer hydrogen bonds between the spin label and the solvent molecules are formed, thereby decreasing the magnitude of the g -tensor shift as well as the hfcc shift. This indirect effect is found to be significantly more important than the direct interaction with the cavity host. Thus, a future strategy to exploit spin labels to study structure; surface properties; and dynamics of proteins, membranes, and other biological complexes in their native environment is to design soluble spin labels containing guest–host complexes with cavities that strike the balance between the protective effect of the cavity, which increases the lifetime of the spin label, and the posing of a minimal effect on the hydrogen bonded water molecules to the R_2NO^\bullet moiety of the spin label. This condition must be met in order to avoid explicit consideration of the encapsulation effect for the EPR data analysis, something that can significantly reduce the information content that can be extracted.

AUTHOR INFORMATION

Corresponding Author

*E-mail: rinkevics@theochem.kth.se.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

This work has been partially funded by the EU Commission (contract INFOS-RI-261523) under the ScalaLife collaboration and has been supported by a grant from the Swedish National Infrastructure Committee (SNIC) for the project “Multiphysics Modeling of Molecular Materials”, SNIC 022/09-25. J.K. thanks The Danish Councils for Independent Research (STENO and Sapere Aude programmes), the Lundbeck Foundation, and the Villum foundation for financial support.

REFERENCES

- (1) Klare, J.; Steinhoff, H.-J. *Photosynth. Res.* **2009**, *102*, 377–390.
- (2) Subczynski, W. K.; Widomska, J.; Feix, J. B. *Free Radical Biol. Med.* **2009**, *46*, 707–718.
- (3) Margittai, M.; Langen, R. *Q. Rev. Biophys.* **2008**, *41*, 265–297.
- (4) Schiemann, O.; Prisner, T. F. *Q. Rev. Biophys.* **2007**, *40*, 1–53.
- (5) Bennati, M.; Prisner, T. F. *Rep. Prog. Phys.* **2005**, *68*, 411–448.
- (6) Mobius, K.; Savitsky, A.; Schnegg, A.; Plato, M.; Fuchs, M. *Phys. Chem. Chem. Phys.* **2005**, *7*, 19–42.
- (7) Subczynski, W. K.; Kusumi, A. *Biochim. Biophys. Acta* **2003**, *1610*, 231–243.
- (8) Prisner, T.; Rohrer, M.; MacMillan, F. *Annu. Rev. Phys. Chem.* **2001**, *52*, 279–313.
- (9) Borbat, P. P.; Costa-Filho, A. J.; Earle, K. A.; Moscicki, J. K.; Freed, J. H. *Science* **2001**, *291*, 266–269.
- (10) Deligiannakis, Y.; Louloudi, M.; Hadjiliadis, N. *Coord. Chem. Rev.* **2000**, *204*, 1–112.
- (11) Hubbell, W. L.; Cafiso, D. S.; Altenbach, C. *Nat. Struct. Biol.* **2000**, *7*, 735–739.
- (12) Hubbell, W. L.; Gross, A.; Langen, R.; Lietzow, M. A. *Curr. Opin. Struct. Biol.* **1998**, *8*, 649–656.
- (13) Berliner, L. *Eur. Biophys. J.* **2010**, *39*, 579–588.
- (14) Abramović, Z.; Brgles, M.; Habjanec, L.; Tomašić, J.; Šentjurc, M.; Frkanec, R. *Int. J. Biol. Macromol.* **2010**, *47*, 396–401.

- (15) Plonka, P. M. *Exp. Dermatol.* **2009**, *18*, 472–484.
- (16) Burks, S. R.; Barth, E. D.; Halpern, H. J.; Rosen, G. M.; Kao, J. P. *Biochim. Biophys. Acta* **2009**, *1788*, 2301–2308.
- (17) Okazaki, S.; Mannan, M. A.; Sawai, K.; Masumizu, T.; Miura, Y.; Takeshita, K. *Free Radical Res.* **2007**, *41*, 1069–1077.
- (18) Matsumoto, K.; Yahiro, T.; Yamada, K.; Utsumi, H. *Magn. Reson. Med.* **2005**, *53*, 1158–1165.
- (19) Bobko, A. A.; Kirilyuk, I. A.; Grigor'ev, I. A.; Zweier, J. L.; Khramtsov, V. V. *Free Radical Biol. Med.* **2007**, *42*, 404–412.
- (20) Bardelang, D.; Banaszak, K.; Karoui, H.; Rockenbauer, A.; Waite, M.; Udachin, K.; Ripmeester, J. A.; Ratcliffe, C. I.; Ouari, O.; Tordo, P. *J. Am. Chem. Soc.* **2009**, *131*, 5402–5404.
- (21) Kirilyuk, I.; Polovyanenko, D.; Semenov, S.; Grigor'ev, I.; Gerasko, O.; Fedin, V.; E., B. *J. Phys. Chem. B* **2010**, *114*, 1719–1728.
- (22) Kirilyuk, I. A.; Bobko, A. A.; Grigor'ev, I. A.; Khramtsov, V. V. *Org. Biomol. Chem.* **2004**, *2*, 1025–1030.
- (23) Krishna, M. C.; Grahame, D. A.; Samuni, A.; Mitchell, J. B.; Russo, A. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 5537–5541.
- (24) Kuppasamy, P.; Li, H.; Ilangovan, G.; Cardounel, A. J.; Zweier, J. L.; Yamada, K.; Krishna, M. C.; Mitchell, J. B. *Cancer Res.* **2002**, *62*, 307–312.
- (25) Swartz, H. M.; Sentjurs, M., II. *Biochim. Biophys. Acta* **1986**, *888*, 82–90.
- (26) Olsen, J. M.; Aidas, K.; Kongsted, J. *J. Chem. Theory Comput.* **2010**, *6*, 3721–3734.
- (27) Olsen, J. M. H.; Kongsted, J.; Sabin, J. R.; Brändas, E. Chapter 3 - *Molecular Properties through Polarizable Embedding*; Academic Press: New York, 2011; Vol. 61, pp 107–143.
- (28) Rinkevicius, Z.; Murugan, N. A.; Kongsted, J.; Aidas, K.; Steindal, A. H.; Ågren, H. *J. Phys. Chem. B* **2011**, *115*, 4350–4358.
- (29) Rinkevicius, Z.; Murugan, N. A.; Kongsted, J.; Frecus, B.; Steindal, A. H.; Ågren, H. *J. Chem. Theory Comput.* **2011**, *7*, 3261–3271.
- (30) Crescenzi, O.; Pavone, M.; De Angelis, F.; Barone, V. *J. Phys. Chem. B* **2004**, *109*, 445–453.
- (31) Barone, V.; Cimino, P.; Pedone, A. *Magn. Reson. Chem.* **2010**, *48*, S11–S22.
- (32) Nose, S. *Mol. Phys.* **1984**, *52*, 255–268.
- (33) Hoover, W. G. *Phys. Rev. A* **1986**, *34*, 2499–2500.
- (34) Parrinello, M.; Rahman, A. *Phys. Rev. Lett.* **1980**, *45*, 1196–1199.
- (35) Case, D. *Amber 8*; University of California: San Francisco, CA, 2004.
- (36) Jorgensen, W. L.; Madura, J. D. *J. Am. Chem. Soc.* **1983**, *105*, 1407–1413.
- (37) Wang, J.; Wolf, R.; Caldwell, J.; Kollman, P.; Case, D. *J. Comput. Chem.* **2004**, *34*, 1157–1174.
- (38) Breneman, C.; Wiberg, K. *J. Comput. Chem.* **1990**, *11*, 361–373.
- (39) Stendardo, E.; Pedone, A.; Cimino, P.; Menziani, M.; Crescenzi, O.; Barone, V. *Phys. Chem. Chem. Phys.* **2010**, *12*, 11697–11709.
- (40) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (41) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.
- (42) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.
- (43) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200–1211.
- (44) Wullen, C. *Die Berechnung magnetischer Eigenschaften unter Berücksichtigung der Elektronkorrelation: Die Multikonfigurations-Verallgemeinerung der IGLO-Methode*. PhD thesis, Ruhr-Universität, Bochum, Germany, 1992.
- (45) Lantto, P.; Vaara, J.; Helgaker, T. *J. Chem. Phys.* **2002**, *117*, 5998–6009.
- (46) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822–8824.
- (47) Perdew, J. P.; Ernzerhof, M.; Burke, K. *J. Chem. Phys.* **1996**, *105*, 9982–9985.
- (48) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- (49) Ernzerhof, M.; Scuseria, G. E. *J. Chem. Phys.* **1999**, *110*, 5029–5036.
- (50) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.
- (51) Gagliardi, L.; Lindh, R.; Karlstrom, G. *J. Chem. Phys.* **2004**, *121*, 4494–4500.
- (52) DALTON. www.daltonprogram.org (accessed Dec. 2011).
- (53) Kawamura, T.; Matsunami, T.; Yonezawa, T. *Bull. Chem. Soc. Jpn.* **1967**, *40*, 1111–1115.
- (54) Ramachandran, C.; Pyter, R. A.; Mukerjee, P. *J. Phys. Chem.* **1982**, *86*, 3198–3205.
- (55) Voinov, M. A.; Ruuge, A.; Reznikov, V. A.; Grigor'ev, I. A.; Smirnov, A. I. *Biochemistry* **2008**, *47*, 5626–5637.
- (56) Owenius, R.; Engström, M.; Lindgren, M.; Huber, M. *J. Phys. Chem. A* **2001**, *105*, 10967–10977.
- (57) Engström, M.; Vaara, J.; Schimmelpfennig, B.; Ågren, H. *J. Phys. Chem. B* **2002**, *106*, 12354–12360.
- (58) D'Amore, M.; Improta, R.; Barone, V. *J. Phys. Chem. A* **2003**, *107*, 6264–6269.
- (59) Rinkevicius, Z.; Telyatnyk, L.; Vahtras, O.; Ruud, K. *J. Chem. Phys.* **2004**, *121*, 5051–5060.
- (60) Improta, R.; Barone, V. *Chem. Rev.* **2004**, *104*, 1231–1254.
- (61) Neugebauer, J.; Louwse, M. J.; Belanzoni, P.; Wesolowski, T. A.; Baerends, E. J. *J. Chem. Phys.* **2005**, *123*, 114101.
- (62) Sinnecker, S.; Rajendran, A.; Klamt, A.; Diedenhofen, M.; Neese, F. *J. Phys. Chem. A* **2006**, *110*, 2235–2245.
- (63) Pavone, M.; Cimino, P.; De Angelis, F.; Barone, V. *J. Am. Chem. Soc.* **2006**, *128*, 4338–4347.
- (64) Barone, V.; Brustolon, M.; Cimino, P.; Polimeno, A.; Zerbetto, M.; Zoleo, A. *J. Am. Chem. Soc.* **2006**, *128*, 15865–15873.
- (65) Carlotto, S.; Cimino, P.; Zerbetto, M.; Franco, L.; Corvaja, C.; Crisma, M.; Formaggio, F.; Toniolo, C.; Polimeno, A.; Barone, V. *J. Am. Chem. Soc.* **2007**, *129*, 11248–11258.
- (66) Pavone, M.; Cimino, P.; Crescenzi, O.; Sillanpää, A.; Barone, V. *J. Phys. Chem. B* **2007**, *111*, 8928–8939.
- (67) Houriez, C.; Ferré, N.; Masella, M.; Siri, D. *J. Chem. Phys.* **2008**, *128*, 244504.
- (68) Houriez, C.; Ferré, N.; Siri, D.; Masella, M. *J. Phys. Chem. B* **2009**, *113*, 15047–15056.
- (69) Houriez, C.; Ferré, N.; Masella, M.; Siri, D. *THEOCHEM* **2009**, *898*, 49–55.
- (70) Pavone, M.; Biczysko, M.; Rega, N.; Barone, V. *J. Phys. Chem. B* **2010**, *114*, 11509–11514.
- (71) Hermosilla, L.; García de la Vega, J. M.; Sieiro, C.; Calle, P. *J. Chem. Theory Comput.* **2011**, *7*, 169–179.
- (72) Ikryannikova, L. N.; Ustyniuk, L. Y.; Tikhonov, A. N. *Magn. Reson. Chem.* **2010**, *48*, 337–349.
- (73) Jayaraj, N.; Porel, M.; Ottaviani, F. M.; Maddipatla, M. V.; Modelli, A.; Da Silva, J. P.; Bhogala, B. R.; Captain, B.; Jockusch, S.; Turro, N. J.; Ramamurthy, V. *Langmuir* **2009**, *25*, 13820–13832.
- (74) Mezzina, E.; Cruciani, F.; Pedullì, G.; Lucarini, M. *Chem.—Eur. J.* **2007**, *13*, 7223–7233.
- (75) Mileo, E.; Mezzina, E.; Grepioni, F.; Pedullì, G. F.; Lucarini, M. *Chem.—Eur. J.* **2009**, *15*, 7859–7862.
- (76) Jockusch, S.; Zeika, O.; Jayaraj, N.; Ramamurthy, V.; Turro, N. J. *J. Phys. Chem. Lett.* **2010**, *1*, 2628–2632.
- (77) Mileo, E.; Yi, S.; Bhattacharya, P.; Kaifer, A. *Angew. Chem., Int. Ed.* **2009**, *48*, 5337–5340.

Performance of Cluster Expansions of Coverage-Dependent Adsorption of Atomic Oxygen on Pt(111)

David J. Schmidt,[†] Wei Chen,[‡] C. Wolverton,[‡] and William F. Schneider^{*,†,§}

[†]Chemical and Biomolecular Engineering, University of Notre Dame, Notre Dame, Indiana 46556, United States

[‡]Material Science and Engineering, Northwestern University, Evanston, Illinois 60208, United States

[§]Chemistry and Biochemistry, University of Notre Dame, Notre Dame, Indiana 46556, United States

 Supporting Information

ABSTRACT: A density functional theory (DFT) database of 66 Pt(111)/O formation energies is presented. We fit this database of formation energies to a range of cluster expansions (CEs) of systematically increasing size and flexibility. We find that the performance of the CE depends upon the property or properties of interest. Pair-wise CEs with up to third nearest neighbor interactions poorly predict all metrics. CEs with five to eight pairwise interactions and one to two triplet interactions predicted formation energies and most ground states accurately but predicted average and differential adsorption energies with modest errors. A larger CE captures average and differential adsorption energies as well as formation energies and ground states. The choice of figures in the CEs is also examined. Pair-wise figures and the linear, 1–1–3, triplet are necessary to obtain CEs that qualitatively reproduce the examined properties; however, other figures are more interchangeable. The electronic and strain components of the adsorbate–adsorbate interactions is studied by comparing a CE of DFT formation energies in which atoms were not allowed to relax to the CEs of the relaxed surface. On an unrelaxed Pt surface, interactions are shorter-ranged interactions and more repulsive at first nearest neighbor separation.

1. INTRODUCTION

Surface adsorption is fundamental to corrosion, gas separations, chromatography, and heterogeneous catalysis. The familiar Langmuir model describes adsorption as binding of adsorbates to chemically unsaturated surface sites, and the Langmuir isotherm describes the limit that these sites are equivalent and independent. For dissociative adsorption of a diatomic adsorbate, the isotherm is given by^{1,2}

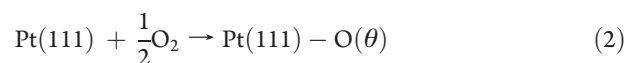


$$K(T) = \frac{\theta_{A*}^2}{P_{A_2} \theta_*^2} \quad (1b)$$

Here, each adsorbate, A, binds with the same adsorption energy at a single surface site of type, *. At equilibrium, the distribution of adsorbates among surface sites is entirely random, to maximize configurational entropy, and the equilibrium constant is written in terms of the adsorbate coverage, θ_{A*} , the vacant site coverage, θ_* , and the pressure of the diatomic adsorbate, P_{A_2} , impinging on the surface. This idealized Langmuir isotherm breaks down when adsorbate–adsorbate interactions become non-negligible.^{2–7} When adsorbate–adsorbate interactions become non-negligible, the adsorption energy of each adsorbate is influenced by its interactions with other adsorbates, and the equilibrium spatial distribution of adsorbates becomes a balance between minimizing these interaction energies and maximizing configurational entropy.

Strong adsorbate–adsorbate interactions are evident in many systems of practical interest. One that we have been particularly interested in due to its relevance to catalytic oxidations,^{8–12} as

well as the oxygen reduction reaction (ORR),¹³ is dissociative O₂ adsorption on the close-packed, hexagonal Pt(111) surface:



The (111) surface exposes face-centered-cubic (fcc) and hexagonal-close-packed (hcp) 3-fold sites, distinguished by the absence or presence of an underlying metal atom (Figure 1). Low energy electron diffraction (LEED), electron energy loss spectroscopy (EELS),¹⁴ nuclear reaction analysis (NRA), and transmission channeling (TC) experiments¹⁵ and supercell density functional theory (DFT) calculations^{11,12,16} agree that at low to moderate coverage O adsorbs preferentially in the fcc hollow sites. Temperature programmed desorption (TPD) and calorimetric experiments,^{17–19} as well as DFT calculations,^{11,12,16,20–22} agree that the interactions between these are primarily repulsive, so that adsorption energies are decreasingly exothermic with increasing coverage. Above approximately $\theta_O = 0.5$ monolayer (ML), adsorbed O begins to populate hcp sites and cause surface reconstructions that further modify adsorption energies.^{16,23}

Lateral interactions between adsorbates can also be manifested in adsorbate orderings. LEED experiments on Pt(111) crystals dosed with oxygen show p(2 × 2)-type ordering at 1/4 and 1/2 ML.^{14,17,18,24} These patterns have been ascribed to a p(2 × 2)-O ordering at 1/4 ML and the three orientations of the p(2 × 1)-O configuration at 1/2 ML.^{18,24} Similar orderings have been observed in scanning tunneling microscopy (STM)

Received: September 19, 2011

Published: December 09, 2011

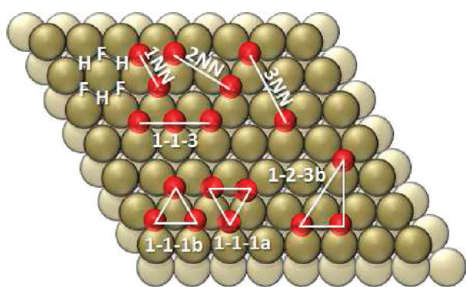


Figure 1. Schematic view of Pt (111) close-packed plane, distinguishing fcc (F) and hcp (H) sites, and illustrating representative cluster interactions.

experiments at 450 K for the $p(2 \times 2)$ -O configuration and 573 K for the $p(2 \times 1)$ -O configuration.²³

DFT calculations can be used to search for the ground states and to model coverage-dependent adsorption at metal surfaces. The Pt(111)–O system has been the subject of a number of such studies,^{11,42,16,20–22,25} recovering the coverage-dependent adsorption, predicting both the $p(2 \times 2)$ -O and $p(2 \times 1)$ -O orderings to be equilibrium ground states.^{12,16,20–22} Due to their computational cost, periodic DFT methods are limited in the sizes of supercells and thus in the range of adsorbate configurations, coverages, and adsorption energies that can be probed. An alternative to direct calculation is to develop a model energy Hamiltonian. In the first-principles cluster expansion (CE) approach,^{26–29} an Ising-type model is fit to the energies of a DFT database of adsorbate configurations. In the CE approach, the energy, E_{CE} , of an adsorbate configuration σ is expanded in polynomial “clusters” or “figures” of the spin variables σ_i :^{21,22,29–34}

$$E_{CE}(\sigma) = N_{\text{sites}}J_e + J_p \sum_i \sigma_i + \sum_{ij} J_{ij}\sigma_i\sigma_j + \sum_{ijk} J_{ijk}\sigma_i\sigma_j\sigma_k + \dots \quad (3)$$

Following the Ising convention, $\sigma_i = +1$ represents the presence and -1 the absence of an adsorbate at site i . A cluster expansion including only the empty, J_e , and point, J_p , effective cluster interactions (ECIs) corresponds to a noninteracting Langmuir model. Interactions between adsorbates are captured in pairwise (J_{ij}), three-body (J_{ijk}), and higher-order terms, where the corresponding sums (eq 3) run over all of the sites. As illustrated in Figure 1, pairwise terms can span first-nearest-neighbor (1NN), second-nearest-neighbor (2NN), etc. separations. Three-body clusters can be linear or triangular of various shapes and sizes, and higher order clusters become increasingly diverse. Equation 3 can be simplified and written as the energy per site times the expected (averaged) value of the spin product (also known as the correlation) of all of the vertices for a given figure across all locations (eq 4):

$$E_{CE}(\sigma)/N_{\text{sites}} = \sum_{\text{figures}} m_a J_a \left\langle \prod_j^{\text{figure}_a} \sigma_j \right\rangle \quad (4)$$

Here, m_a is the multiplicity (number of symmetry equivalent rotations and reflections) of the figure. The unknown ECI can be fit to the DFT energy/configuration database using a least-squares algorithm. The infinite basis set of figures is complete

and orthogonal so that the expansion is exact in that limit.³⁵ For a finite-sized database, the practical challenge is to choose a compact set of figures that represents the energies reliably without introducing artifacts of overfitting. Once parametrized, the cluster expansion can be used to rapidly calculate the energy of any arbitrary adsorbate configuration.

Several CEs have been reported for the fcc Pt(111)–O system^{20–22} based on fittings to the DFT energies of a relatively small number (15 to 16) of O configurations. Each of these CEs captures the $p(2 \times 2)$ -O and $p(2 \times 1)$ -O ground state configuration but differ in their predictions of other ground states. While the CEs are typically fit to formation energies relative to the clean Pt(111) surface and oxygen (either as a full ML on the Pt(111) surface or as O₂ in a vacuum), a common application of the CE is the calculation of energy differences corresponding to the addition or removal of a single adsorbate. Such models have been used in Monte Carlo simulations to predict surface adsorbate phase diagrams^{20,21} and recently to model coverage-dependent kinetic phenomena.³⁶ To date, however, a detailed comparison of the performance of the CEs for formation and differential adsorption energies has not been carried out.

Here, we report an extensive DFT database of 66 Pt(111)–O configurations in supercells ranging from 1 to 16 unique adsorption sites. We fit this database of formation energies to a range of CEs of systematically increasing size and flexibility. We find that the performance of the CE depends upon the property or properties of interest. More specifically, ground states and formation energies can be predicted reliably with moderate-sized CEs, but differential adsorption energies require larger CEs to achieve the same accuracy. The choice of figures in cluster expansions is also examined. It is found that many figures are interchangeable but that short-ranged pairwise figures as well as the linear 1–1–3 triplet (Figure 1) are necessary to obtain CEs that correctly capture the qualitative features of the Pt(111)/O system. Additionally, strain effects are studied by isolating the electronic effects in a CE of a database of formation energies for unrelaxed configurations. Interactions are shorter ranged and more repulsive at a 1NN separation in the unrelaxed configurations; adsorbate-induced surface strain is responsible for longer-ranged repulsions and contributes a 1NN attractive interaction that decreases, but does not overcome, the intrinsic electronic 1NN repulsion.

2. COMPUTATIONAL METHODS

We used the Vienna ab initio package (VASP)^{37–40} to perform plane-wave, supercell DFT calculations within the PW91 implementation of the generalized gradient approximation (GGA)⁴¹ and a projector augmented wave (PAW) treatment of core electronic states.^{42,37,39,40} Plane waves were included in the DFT calculations to an energy cutoff of 400 eV. Bulk Pt energy calculations were performed on unit face centered cubic (fcc) cells with a well-converged $30 \times 30 \times 30$ Γ -centered k -point mesh. The lattice constants of these cells were uniformly distributed near the experimental value of 3.912 Å.⁴³ Fitting these energies to the Birch–Murnaghan equation of state^{44,45} yields a computed lattice constant of 3.986 Å, which is used to determine the dimensions of supercells in the subsequent slab calculations. Pt(111) surface calculations were performed using a slab model consisting of four Pt layers, an atomic oxygen adsorbate layer, and four vacuum layers (Figure 2). The bottom Pt layer was fixed, and

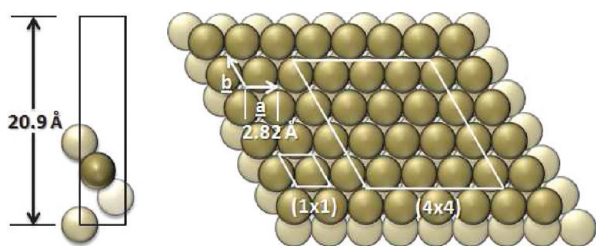


Figure 2. Side and top views of the smallest (1 × 1) periodic cell as well as the top view of the largest (4 × 4) periodic cell and unit vectors.

the remaining layers were relaxed. For comparison, we calculated DFT formation energies of the p(2 × 1)-O and full ML configurations using seven Pt layers and the bottom layer fixed, and with four Pt layers and the bottom two layers fixed. Formation energies differed by less than 4 meV/O. A Γ -centered k -point mesh with at least 85 k points per \AA^{-3} was used. O atoms were displaced from high symmetry positions and relaxed until the energy difference between subsequent optimization steps was less than 0.1 meV. Single point (no ionic relaxations) energy calculations were done using the tetrahedron method with Blöchl corrections⁴⁶ to determine final relaxed GGA energies. The effect of dipole corrections was tested on the clean, p(2 × 1)-O, and full ML configurations using a compensating dipole sheet at the center of the vacuum and parallel to the surface, as implemented in the VASP code.⁴⁰ Formation energy differences were only 1.0 and 2.5 meV/site for the p(2 × 1)-O and full ML surfaces, respectively, reflecting the small dipole of the Pt–O bond. Fitted energies do not include dipole corrections.

We calculated the bond length and GGA energy of triplet molecular oxygen in a 20 × 20 × 20 supercell to be 1.225 Å and −9.8 eV/O₂, respectively. The harmonic vibrational frequency, calculated by finite difference on the forces using displacements of 0.01 Å, was calculated to be 1550 cm^{−1}. Both the bond length and frequency are close to the experimental values of 1.208 Å⁴⁷ and 1556 cm^{−1},⁴⁸ respectively. For comparison, the molecular O₂ energy was also calculated from the experimental heat of formation of gaseous water extrapolated to 0 K ($H_{0,\text{H}_2\text{O}}^{\text{F}}$),⁴⁹ and the zero-point corrected GGA energies of H₂O and H₂ ($E_{0,i}$) were calculated in 20 × 20 × 20 supercells with an energy cutoff of 700 eV:

$$E_{0,\text{O}_2} = E_{0,\text{H}_2\text{O}} - H_{0,\text{H}_2\text{O}}^{\text{F}} + E_{0,\text{H}_2} \quad (5)$$

The H₂ vibrational frequency is 4161 cm^{−1}⁵⁰ and those of water are 3657, 1595, and 3756 cm^{−1}.⁴⁹ The O₂ energy calculated this way is −9.5 eV/O₂. The difference between the two references of 0.2 eV/O (0.3 eV/O₂) reflects the intrinsic uncertainty in the GGA oxygen energy. We use the O₂ reference in the results reported here. The opposite choice would uniformly shift all energies by 0.2 eV/O. Relative errors for atomic oxygen on the Pt(111) surface are expected to be significantly less as GGA DFT energies are able to correctly predict the phase behavior.^{12,20–22} The error due to the uncertainty in the oxygen reference only affects the empty and point figures in any CE containing both of these figures, such that

$$\delta J_e = \frac{1}{4} \delta E_{\text{O}_2} \quad (6)$$

$$\delta J_p = -\frac{1}{4} \delta E_{\text{O}_2} \quad (7)$$

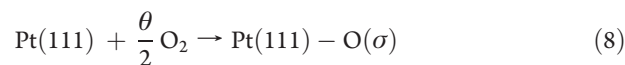
δJ_e and δJ_p are the uncertainties in the empty and point ECIs due to the uncertainty in the oxygen reference, δE_{O_2} . This relation can be derived by setting $(\theta/2)\delta E_{\text{O}_2}$ equal to the cluster expansion Hamiltonian (eq 3). Pair and higher order terms are not affected by any DFT error in E_{O_2} .

A series of cluster expansions of the adsorbate formation energies was constructed from the DFT oxygen energy database (see eq 9 in section 3.1) by considering up to 10th nearest neighbor (10NN) pairs (the largest supercell was 11.3 Å by 11.3 Å), triplets with a single side up to 5NN separation, and larger figures with up to six sites and no two sites separated by more than a 3NN distance. This yielded a candidate set of 10 pair, 29 triplet, 10 quadruplet, 7 quintuplet, and 4 sextuplet figures. Multiple steepest descent sweeps, with various sets of starting figures, were used to search for cluster expansions with the smallest errors. Tools from the Alloy Theoretic Automated Toolkit (ATAT) were used to calculate cross-validation (cv) scores for the steepest descent.^{26,27,51}

3. RESULTS

3.1. DFT Database. The structures and DFT energies of a total of 66 relaxed configurations of fcc atomic oxygen on Pt(111) were computed at coverages between 0 and 1 ML. These configurations were selected iteratively, starting from the smallest ordered structures and adding more complicated ones based on predictions from the CEs.^{26,27,51} In all cases, adsorbed O remains in fcc sites during optimization and the surface Pt remains near the same surface plane. Configurations and energies are included in the Supporting Information.

To compare the stabilities of these various configurations, we first calculate the zero-point corrected DFT formation energy, $E_f(\sigma)$, per site relative to the clean surface and molecular O₂:



$$E_f(\sigma) = \frac{E_0(\sigma) - E_0(0)}{N} - \frac{\theta}{2} E_{0,\text{O}_2} \quad (9)$$

$E_0(\sigma)$ is the zero-point corrected DFT energy of configuration σ . $E_0(0)$ is the energy of the clean surface calculated in an identical supercell. N is the number of fcc surface sites in the supercell. $\theta = N_{\text{O}}/N$ is the oxygen coverage of the configuration, and E_{0,O_2} is the zero-point corrected DFT energy of O₂. We neglect the minor changes in the Pt vibrational spectrum with adsorption and include vibrational contributions of adsorbed O, assuming these to be independent of coverage. DFT-calculated vibrational frequencies for an adsorbed O in a p(4 × 4) supercell are 429, 380, and 377 cm^{−1},¹² similar to the experimental values of 480 and 400 (doubly degenerate) cm^{−1}.¹⁴ The zero-point correction is thus linear in coverage and contributes 0.15 eV/O. Even with the dense k -point sampling used here, we find that it is important to construct energy differences between identically sized supercells to maximize error cancellation; otherwise, slight variations in the per site $E_0(0)$ with supercell size and shape lead to spurious predictions of relative energies and of ground states that propagate into errors in the CE fitting.

Figure 3a plots the GGA-computed formation energies (eq 9) vs O coverage. The convex hull connecting the stable ground-state configurations is indicated with a line, and configurations along the hull are sketched in Figure 4, including the corresponding periodic supercell. The negative initial slope of the hull reflects

exothermic dissociative O₂ adsorption at low coverage, and the upward curvature reflects the repulsive interactions between adsorbates. Notable are the numerous configurations near but not breaking the convex hull.¹² The large number of configurations

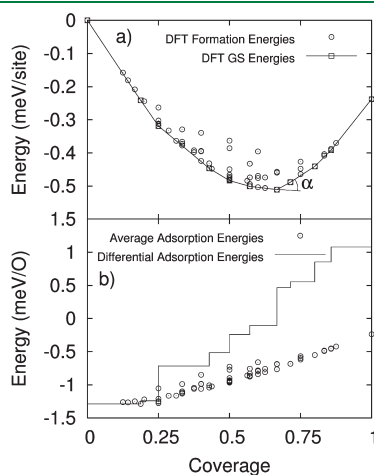


Figure 3. (a) DFT formation energies ($E_F(\sigma)$) for the fcc O/vacancy system on Pt(111). Ground-state configurations are highlighted and connected by lines. The definition of the hull exterior angle, α , is indicated. (b) DFT differential ($\epsilon_{\text{Ads}}(\sigma_1, \sigma_2)$) and average ($E_{\text{Ads}}(\sigma)$) adsorption energies.

close to or on the convex hull suggests, at practically relevant temperatures and coverages away from the strong ground states, that no one configuration will dominate the surface and, instead, regions of coexistence and/or disorder will dominate.

Table 1 summarizes the coverages and symmetries (following the 1996 IUPAC convention)⁵² of the computed ground states. To quantify the extent to which the various ground states break the convex hull, we calculate the hull exterior angle α , defined as the exterior angle minus 180° (Figure 3a). Prominent ground states with large exterior angles include the experimentally observed $p(2 \times 2)$ -O ordering (Figure 4c), in which all O atoms are equivalent and third-nearest neighbor to one another, and $p(2 \times 1)$ -O ordering (Figure 4e), in which O atoms form first-nearest-neighbor rows separated by empty rows.¹² The $p(\sqrt{3} \times \sqrt{3})$ -2O ground state (Figure 4g) has a hexagonal symmetry and the largest hull exterior angle of all ground states. The $3/7$ and $4/7$ ML configurations (Figure 4d and f) have α values close to that of the $p(2 \times 1)$ -O one and present similar rows of first-nearest-neighbor separated O. In the $3/7$ ML configuration, rows of length three are offset by the $p(100)$ vector, so that the row ends are at second-nearest-neighbor separations. In the $4/7$ ML configuration, these rows are of length four, offset by the $p(110)$ vector, so that all row ends are at first-nearest-neighbor separations. These two configurations are interesting in that not all adsorbed O atoms are equivalent. A similar structure at $2/5$ ML containing $p(2 \times 1)$ -O rows of length two was identified as a ground state in a previous DFT study.²¹ We calculate the

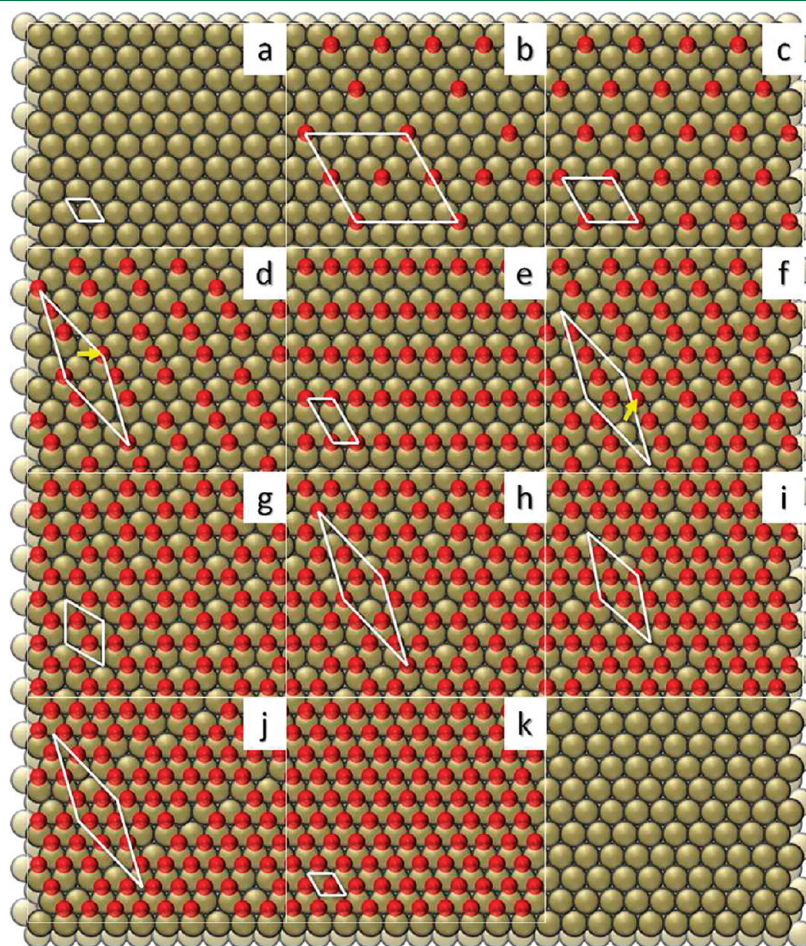


Figure 4. Ground state configurations and periodic supercells in the fcc O/vacancy system on Pt(111). See Table 1 for symmetries.

Table 1. DFT-Computed Ground State Configurations and Hull Exterior Angles, α

	coverage (ML)	α (deg)	symmetry
(b)	3/16	3	$p(4 \times 4)$ -3O
(c)	1/4	28	$p(2 \times 2)$ -O
(d)	3/7	11	$\begin{pmatrix} 1 & 3 \\ -1 & 4 \end{pmatrix}$ -3O
(e)	1/2	15	$p(2 \times 1)$ -O
(f)	4/7	8	$\begin{pmatrix} 1 & 3 \\ -1 & 4 \end{pmatrix}$ -4O
(g)	2/3	30	$(\sqrt{3} \times \sqrt{3})R30^\circ$ -2O
(h)	5/7	5	$\begin{pmatrix} 1 & 3 \\ -1 & 4 \end{pmatrix}$ -5O
(i)	4/5	17	$\begin{pmatrix} 1 & 3 \\ -1 & 4 \end{pmatrix}$ -4O
(j)	6/7	13	$\begin{pmatrix} 1 & 3 \\ -1 & 4 \end{pmatrix}$ -6O

2/5 ML structure to lie just above the computed convex hull. Because of the similarity between all of these structures and the $p(2 \times 1)$ -O configuration, it may be difficult to distinguish these in LEED or STM experiments; nonetheless, their appearance has implications for surface adsorption energies.

The average adsorption energy, $E_{\text{ads}}(\sigma)$, per oxygen is related to the formation energy by the coverage, θ :

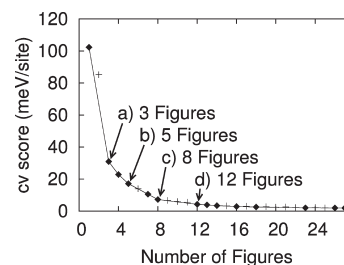
$$E_{\text{ads}}(\sigma) = \frac{E_{\text{F}}(\sigma)}{\theta} \quad (10)$$

Figure 3b plots the GGA-computed E_{ads} vs coverage. The low coverage limiting adsorption energy is computed to be -1.3 eV per O using the GGA O_2 reference (-1.5 eV per O using the GGA $\text{H}_2\text{O}/\text{H}_2$ reference), which can be compared to the -1.1 eV/O inferred from an analysis of oxygen TPD.^{17,18,53} Calorimetric measurements yield a more exothermic value of -1.6 eV/O.¹⁹ Above 1/4 ML, the average energy rises gradually, again reflecting accumulated repulsive interactions between adsorbates. The differential adsorption energy (ε_{ads}) between subsequent ground states σ_1 and σ_2 is the slope of the connecting convex hull, or equivalently the difference in formation energy divided by the difference in coverage:

$$\varepsilon_{\text{ads}}(\sigma_1, \sigma_2) = \frac{E_{\text{F}}(\sigma_1) - E_{\text{F}}(\sigma_2)}{\theta(\sigma_1) - \theta(\sigma_2)} \quad (11)$$

GGA-calculated differential adsorption energies are plotted as the staircase in Figure 3b. ε_{ads} is equivalent to E_{ads} at low coverage but exhibits discontinuous jumps at the ground states of heights related to the hull exterior angle. The differential adsorption energy becomes significantly positive at 2/3 ML coverage; even with a generous estimate of the GGA error, O_2 dissociation into fcc sites is predicted to become endothermic above this 2/3 ML coverage.

3.2. Cluster Expansions. We next fit the oxygen adsorbate formation energy DFT database to a set of cluster expansion models. While in principle the infinite cluster expansion is exact, in practice the expansion must be truncated at some finite number

**Figure 5.** Cross validation (meV/site) score versus number of figures.

of terms, and the selection of an optimal set of fitting figures is one of the primary challenges to the CE approach.^{22,35} The most common measure of the predictive capability of a CE is the leave-one-out cross-validation, or CV score.^{20–22,26–29,32} For a given set of basis figures, the effective cluster interactions (ECIs) are determined by least-squares fitting to all but one of the DFT formation energies and the prediction error in the one excluded configuration determined. This predicted error is calculated for every configuration in the DFT database and the root-mean-square (RMS) value of all errors is the CV score:²⁶

$$\text{CV} = \sqrt{\langle (E_{\text{s}}^{\text{CE}}(\sigma) - E_{\text{F}}(\sigma))^2 \rangle} \quad (12)$$

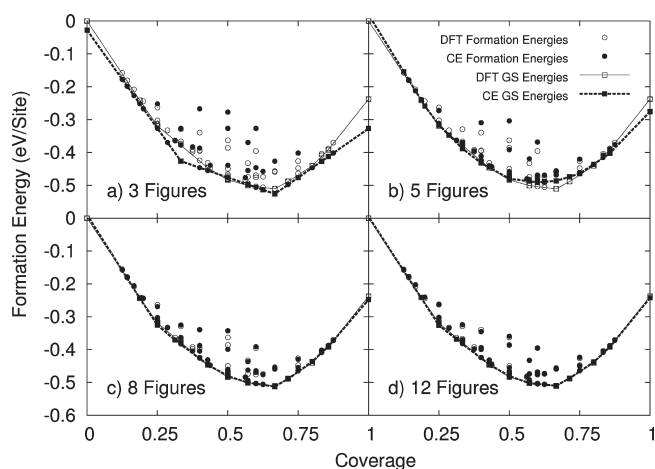
Here, $E_{\text{s}}^{\text{CE}}(\sigma)$ is the predicted energy per site of configuration σ , excluding that configuration from the fit, and $E_{\text{F}}(\sigma)$ is the corresponding DFT-calculated formation energy. Our candidate set of figures included pair, triplet, quadruplet, quintuplet, and sextuplet figures as discussed in section 2. We built up successively larger cluster expansions using steepest descent searches in which each subsequent addition or removal of a figure is done by searching over all possible candidates, adding or removing the figure that maximally decreases the CV score. To overcome local minima in this search procedure, the addition and/or removal of two figures simultaneously was performed when this resulted in a lower CV score. The result of this search is shown in Figure 5, plotted as CV score vs number of symmetry-distinct figures. Analogous to the formation energies, a “convex hull” of CEs can be identified, where kinks represent significant changes in the efficiency of the CE for predicting formation energies.

The lowest cv score did not always monotonically decrease as the number of configurations in the DFT database grew. The cv score typically decreased on the order of 1 meV/site but occasionally increased significantly as new oxygen configurations expanded configuration space. The cv score did monotonically decrease after the database reached 25 oxygen configurations, suggesting that the relevant configuration space had been covered. The iterative DFT and CE procedure provides a valuable self-consistency check: we were able to identify several miscalculated DFT energies by the inability to fit the energies and associated configurations within a CE.

The CE representation of the Langmuir model (eq 1) would include only empty and point terms and, given the relatively strong adsorbate interactions in the Pt(111)/O system, such a model has a very large cv score of 85 meV/site. Table 2 shows the computed ECIs for a CE that adds the 1NN figure (Figure 1), the CE at the first kink in Figure 5. (For counting purposes, we term this a “3 figure” CE, identifying the empty and point as the first and second figures). This 1NN pairwise term has a large positive ECI that strongly disfavors 1NN O pairs and improves the cv score to 31 meV/site. A 3 figure CE with empty, point, and 1NN

Table 2. Effective Cluster Interactions (ECI) for Four Cluster Expansions

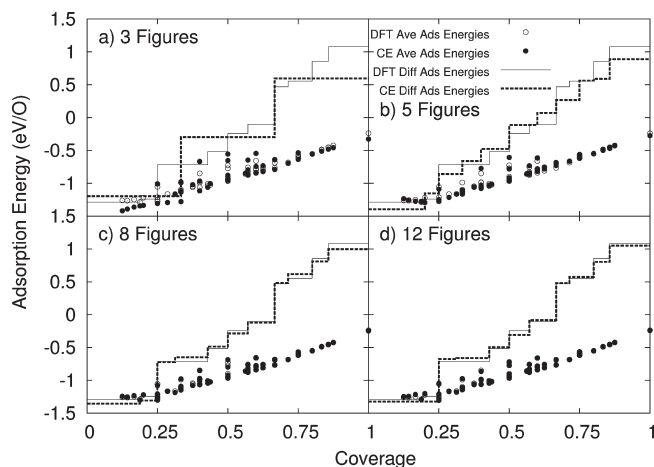
figure	ECI (meV/site)			
	3 figure CE	5 figure CE	8 figure CE	12 figure CE
empty	-402	-424	-423	-426
point	-150	-148	-153	-159
1NN	75	65	55	53
2NN		21	14	12
3NN		21	11	7
4NN			7	7
5NN			7	8
6NN				3
7NN				2
8NN				3
1-1-3			8	7
1-2-3b				2

**Figure 6.** Cluster expansion fitted and DFT formation energies of four CEs of increasing size.

figures can only predict ground states at 0, 1/3, 2/3, and 1 ML coverage and therefore cannot predict the proper phase behavior for dissociative oxygen adsorption on Pt(111).⁵⁴

The 2NN and then 3NN pairs are the two next most important figures found in the CE search, and CEs adding these two appear as kinks at 23 and 17 meV/site in Figure 5, respectively. These terms also have positive ECIs that again disfavor O at these separations. While these two-body terms are the most important O–O interactions,¹² pairwise figures alone are not able to reproduce the asymmetry in the ground state configurations about 1/2 ML. Figure 6a contrasts the DFT-computed formation energies with those predicted by the 3-figure CE. This 3-figure CE fails to reproduce the ground state at 1/4 ML coverage, predicts a spurious ground state at 1/3 ML, predicts incorrect ground state oxygen configurations at 3/7 and 4/7 ML, and errs seriously in the energies of many of the nonground-state configurations. The 5-figure CE in Figure 6b improves the energy predictions significantly in most cases and predicts correct ground states at 3/7 and 4/7 ML but produces unphysical ground states at 1/3 ML and 3/4 ML.

These difficulties are reflected in the CE-computed average and differential adsorption energies, shown in Figure 7. Average

**Figure 7.** Cluster expansion fitted and DFT, average, and differential adsorption energies plotted against coverage for 3 figure CE, 5 figure CE, 8 figure CE, and 12 figure CE.

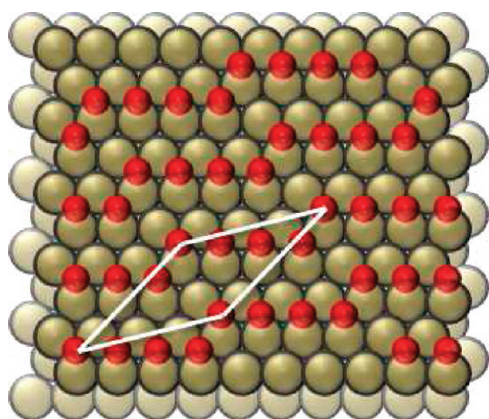
adsorption energies are captured within 0.19 eV/O for most configurations even in the 3 figure CE; the 5 figure CE reduces the largest errors to 0.13 eV/O. In contrast, because of errors in the ground state predictions, differential adsorption energies are reproduced much less well, with errors up to 0.5 eV/O at some coverages in the 3 figure CE and 0.3 eV/O in the 5 figure CE. In general, pairwise only models are not able to quantitatively reproduce coverage-dependent energies or equilibrium configurations of O on Pt(111).

The next minor kink in the *cv* vs figure search is at a seven figure CE that adds the 1-1-3 linear triplet and the fourth-nearest-neighbor (4NN) pair figure (Figure 1), decreasing the *cv* score to 10.5 meV/site. Adding the fifth-NN pair produces a strong kink in Figure 5 at an eight-figure CE with a CV score of 7.2 meV/site. The figures in this 8-figure CE are identical to the Pt(111)–O figures identified in another report.²⁰ As shown in Figure 6c, this eight-figure CE captures the energies of nearly all of the configurations near the ground state hull and deviates substantially only for a few configurations far from the hull. The worst fit configurations are striped ones with large numbers of 1NN O–O and vacant site–vacant site pairs. As shown in Figure 7, this 8-figure CE captures average adsorption energies within 40 meV/O for all configurations. The locations of the steps in the differential formation energies are identified correctly, and differential adsorption energy errors are less than 84 meV/O.

It is interesting to consider the role of the 1-1-3 linear triplet in the Pt(111)/O system. A triplet or larger odd-bodied figure is necessary for introducing asymmetry into the ground state predictions. In introducing the asymmetry, the 1-1-3 linear triplet helps capture the prominent $p(\sqrt{3} \times \sqrt{3})\text{-}2\text{O}$ and $p(2 \times 2)\text{-O}$ configurations without disfavoring the $p(2 \times 1)\text{-O}$ ground state configuration. The 1-1-3 correlations (average spin product, see eq 4) are -1 , $-1/2$, and 0 for the $p(\sqrt{3} \times \sqrt{3})\text{-}2\text{O}$, $p(2 \times 2)\text{-O}$, and $p(2 \times 1)\text{-O}$ configurations, respectively, allowing a positive ECI value to favor the $p(2 \times 2)\text{-O}$ ground state over the $p(2 \times 2)\text{-}3\text{O}$ configuration (their correlations are necessarily opposite for odd-bodied figures) and the $p(\sqrt{3} \times \sqrt{3})\text{-}2\text{O}$ ground state over the $p(\sqrt{3} \times \sqrt{3})\text{-O}$ configuration, see Table 3. Any triplet must favor either the 3/7 or 4/7 ML ground state and disfavor the other for any nonzero correlation,

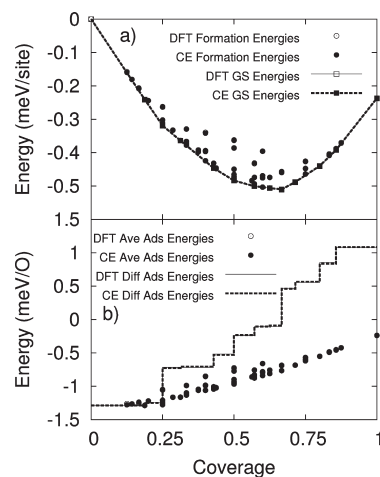
Table 3. Figure Correlations for the Most Prominent Ground States and Their Spin-Flipped Conjugates

σ or θ	figure/correlation				
	1-1-3	1-1-1a	1-1-1b	1-1-2	1-2-3b
$p(2 \times 2)$ -O	-0.5	0.5	0.5	0.5	-0.5
$p(\sqrt{3} \times \sqrt{3})$ -O	1	1	1	-0.33	-0.33
3/7	0.24	0.43	0.43	-0.14	0.05
$p(2 \times 1)$ -O	0	0	0	0	0
4/7	-0.24	-0.43	-0.43	0.14	-0.05
$p(\sqrt{3} \times \sqrt{3})$ -2O	-1	-1	-1	0.33	0.33
$p(2 \times 2)$ -3O	0.5	-0.5	-0.5	-0.5	0.5

**Figure 8.** The 12-figure CE predicted ground state at 1/2 ML O.

as their correlations are necessarily opposite. The 1-1-3 triplet has a correlation of ± 0.24 for these ground states, relatively small compared to the correlations of the $p(2 \times 2)$ -O and $p(\sqrt{3} \times \sqrt{3})$ -2O ground states. Of the three previously published Pt(111)-O cluster expansions, all identified this linear 1-1-3 triplet as an important term.²⁰⁻²² It is interesting to note that the most compact triplets, 1-1-1a and 1-1-1b in Figure 1, which differ as to whether they encompass an fcc or hcp hollow, do not appear in any of the CEs constructed in this work. The correlation for $p(\sqrt{3} \times \sqrt{3})$ -2O is -1 for both 1-1-1 triplets, while the correlation for $p(2 \times 2)$ -O is 1/2 for both triplets. Because these are opposite in sign, any nonzero value of the ECI disfavors one or the other of these two ground states. Similarly, these triplets have moderate correlations of $\pm 3/7$ for the ground states at 3/7 and 4/7 ML.

Four more figures are added to reach the next significant kink in Figure 5, including the 6NN to 8NN pairs along with either the 1-2-3b or 1-1-2 triplet, depending on the exact figure selection procedure. The ECIs for all of these additional figures are 3 meV or less. As shown in Figure 6d, the 12-figure CE captures the formation energies of essentially all 66 configurations within the uncertainty of the DFT model, with greatest improvements in formation energy for configurations that are furthest from the hull. This 12-figure CE captures the high-energy, striped configurations much better than the 8-figure CE. These improvements come at a cost, however: this 12-figure CE predicts the configuration shown in Figure 8 to be 2.5 meV/site lower in energy than the $p(2 \times 1)$ -O one, opposite the 3.4 meV/site higher predicted by the DFT. As shown in Figure 7, the 12-figure

**Figure 9.** DFT and 27-figure cluster expansion (a) formation energies and (b) adsorption energies.

CE reproduces average adsorption energies essentially quantitatively (up to 0.03 eV/O error) and differential formation energies with a maximum error of 0.08 eV/O.

Neither 1-2-3b nor 1-1-2 triplet promote the prominent $p(2 \times 2)$ -O, $p(2 \times 1)$ -O, and $p(\sqrt{3} \times \sqrt{3})$ -2O ground state structures. The 1-2-3b triplet correlations have opposite signs for the $p(2 \times 2)$ -O and $p(\sqrt{3} \times \sqrt{3})$ -2O ground states and therefore cannot promote both ground states, and the 1-1-2 triplet is fit such that the $p(2 \times 2)$ -O and $p(\sqrt{3} \times \sqrt{3})$ -2O ground states are less energetically favorable. Both triplets have correlations of 0 for the $p(2 \times 1)$ -O configuration. As these triplets do not favor the ground states, the main contribution of these triplets appears to be in better fitting with the high energy, striped configurations.

Successively adding figures beyond the 12-figure expansion produces a series of CEs with slowly decreasing CV scores. No qualitatively significant new figures appear in this search. The largest CE constructed here has 27 figures and a CV score of 1.9 meV per site. The ECI for this 27-figure CE are included in the Supporting Information. Figure 9a compares computed and predicted formation energies and Figure 9b the average and differential adsorption energies vs O coverage. This 27-figure CE reproduces all quantities essentially quantitatively (maximum error of 0.02 eV/O). The power of the CE is in its ability to quickly predict energies for adsorbate configurations too numerous or large to practically calculate with DFT. To illustrate this predictive ability, 20 468 O configurations with up to 15 surface sites were generated using the ATAT and formation, average adsorption, and differential adsorption energies predicted. The results are shown in Figure 10. The general shape of the formation energy hull is preserved, but as readily seen, there are very many configurations that appear near the hull. These configurations correspond to various defects in the perfectly ordered configurations. New ground states do appear on the convex hull; however, these ground states have α angles less than 10° and break the convex hull by only a few millielectronvolts.

3.3. Surface Strain Effects. The effects of surface strain on adsorbate binding energies have been investigated experimentally and with first principle approaches.⁵⁵⁻⁵⁸ Surfaces under expansion bind adsorbates more exothermally,^{55,56} while surfaces under compression bind adsorbate less exothermally.⁵⁶ It has also been seen that adsorbates, including atomic oxygen,

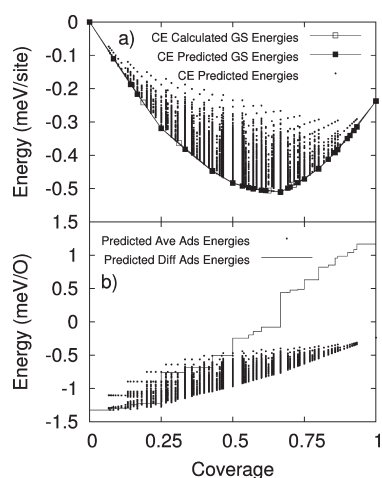


Figure 10. The 27-figure CE predicted (a) formation energies and (b) adsorption energies. No new ground states predicted with $\alpha > 10^\circ$.

induce strain in surfaces;^{12,59} however, the role of strain on adsorbate–adsorbate interactions has not been reported. To examine the effects of strain in adsorbate–adsorbate interactions, we evaluated the energies for a smaller database of atomic oxygen adsorbate configurations in which the platinum atoms are fixed at their clean surface positions and oxygen atoms are fixed in their full ML positions. (Separate tests showed that fixing or relaxing the O atoms had a small effect on the energies and no consequences for the conclusions here.)

The GGA formation energies for the fixed system are shown in Figure 11. The formation energies of the fixed system are similar to the relaxed system in that dissociative O₂ adsorption is exothermic at low coverage, the interactions are generally repulsive, as seen by the upward curvature of the convex hull, and ground states with large α angles exist at 1/4, 1/2, and 2/3 ML coverage. The formation energies are different in that O₂ adsorption is less exothermic than the relaxed system, ground states with large α angles are predicted at 1/3 and 3/4 ML coverage, minor ground states have appeared and disappeared at various other coverages, and there is generally more symmetry about 1/2 ML coverage.

We fitted these fixed Pt GGA energies to a CE following the procedure above. A CE with only four figures (empty, point, 1NN, and 2NN) is able to capture the fixed Pt formation energies with a CV score of 12 meV/site. In contrast, the same-sized CE fits the relaxed surface energies much less well, yielding a CV score of 23 meV/site. Table 4 shows an optimized 19-figure CE of the fixed Pt energies with a CV score of 1.8 meV/site. From comparison with Table 2, both types of figures and the strengths of the corresponding ECIs change. The ECI of the 1NN pair interaction is 16 meV more repulsive in the fixed system than when the Pt's are allowed to relax; however, all other pairwise ECIs are significantly smaller. For example, the ECI of the 2NN pair dropped to 3 meV from 8 meV, and interestingly, the 3NN pair interaction goes from being a repulsive 7 meV to an attractive -4 meV. The optimized relaxed and fixed system CEs also have different multibodied figures. The 1–1–3 triplet, which is most important for the relaxed system, does not appear as an important figure in any of fixed Pt CEs we examined. Oddly, the multibodied figure that has the largest impact on the CV score is the most compact quintuplet, identified in Table 2 as 5(1)–(2) for the five 1NN and one 2NN separations in the figure. The fit of almost every oxygen configuration moderately improves with

Table 4. Effective Cluster Interactions (ECI) for Fixed System CE

figure	ECI (meV)
empty	-310
point	-117
1NN	66
2NN	3
3NN	-4
4NN	6
5NN	1
8NN	1
1–1–1b	-2
2–2–2	4
3–3–3a	1
1–4–4	1
5–5–5a	1
5(1)–(2)	2
2(1)–2(2)–2(3)	1
2(1)–2(2)–2(3)	1
7(1)–2(2)–(3)	2
4(1)–4(2)–2(3)	-2
5(1)–2(2)–3(3)	1

this quintuplet, the most noticeable improvements being for the ground state configurations at 1/3 and 2/3 ML coverage. The remaining multibodied figures fall into at least one of two categories, equilateral (or isosceles in one case) triangular figures or short-ranged figures with up to 3NN pairwise separations, but mostly 1NN and 2NN separations. Interactions are generally shifted from relatively long-ranged interactions, in the relaxed system, to short-ranged interactions in the fixed system.

The Pt–O bond has a modest dipole moment,⁶⁰ so that electrostatics contribute little to the interactions. The short range of the interactions in the fixed Pt case can be understood in terms of simple chemical unsaturation arguments.¹ When an atomic oxygen adsorbate is on the surface, it bonds with the nearest platinum atoms, quenching some of their chemical unsaturation. Because the nearest platinum atoms are less chemically unsaturated, other atoms (including 1NN adsorbates) bind to these nearest platinum atoms less exothermically. A second adsorbate, binding 1NN with respect to the first adsorbate, will share a Pt with the first adsorbate and bind less favorably than the first adsorbate; hence there is an effective 1NN repulsion. The next nearest platinum atoms are now more chemically unsaturated (since they are not bound as tightly to the nearest platinum atoms) and will bind more strongly to other atoms than they did without an adsorbate on the surface. However, each nearest Pt atom has nine neighboring Pt atoms (including subsurface atoms) compared to only three nearest platinum atoms, so the effect is much less pronounced for the next nearest Pt atoms. As changes in the chemical saturation are propagated outwardly in three dimensions, they rapidly decay such that adsorbate interactions, due to chemical saturation/unsaturation, are short-ranged.

We performed fixed and relaxed calculations on a 6×6 supercell of a single O at an FCC site. Relaxation of the surface lowers the energy by over 200 meV/O. Figure 12 shows that Pt's 1NN to O are uniformly displaced laterally,¹² causing a local expansion of the Pt–Pt bonds (shown with red triangles), while those farther away are compressed (shown with blue triangles).

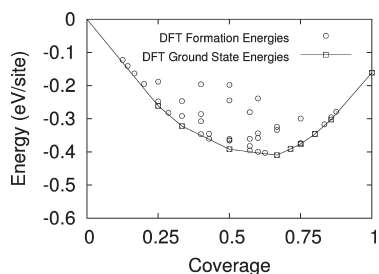


Figure 11. DFT formation energies for the fcc O/vacancy system on Pt(111) with platinum atoms fixed in clean surface locations. Ground-state configurations are highlighted and connected by lines.

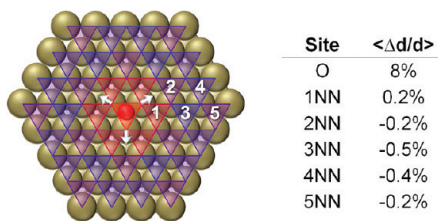


Figure 12. Single oxygen adsorbate on $p(6 \times 6)$ surface. Strain increases average Pt–Pt bond length locally (red triangles) and decreases remote average Pt–Pt bond lengths (blue triangles). The percent change of average Pt–Pt bond lengths around each type of site is shown in the inset table.

The inset table shows the average change in the Pt–Pt bond lengths around the adsorbed O and around each potential adsorption site up to 5NN from the adsorbed O. The O adsorbate induced strain increases the average Pt–Pt bond length by 8% for the Pt atoms around the O adsorbate and by 0.2% for the Pt atoms around the potential 1NN adsorption sites while decreasing the average Pt–Pt bond length by 0.2 to 0.5% for Pt atoms around the other potential adsorption sites. As strain is a long-range effect,^{61–65} strain-induced relaxations persist over longer distances than the chemical saturation/unsaturation effects.

As noted above, surfaces under compressive strain bind adsorbates less strongly, while those under expansion bind adsorbates more strongly.⁵⁶ Therefore, adsorbate-induced strain is attractive at 1NN and repulsive at longer distances. Comparing Tables 2 and 4 confirms this trend. The long-range, strain-induced repulsions increase the relative probability of adsorbates binding at 1NN, acting as a pseudoattraction.

4. CONCLUSIONS

DFT-fitted cluster expansions provide a general means of incorporating coverage dependence into models of surface adsorption. Here, we have examined the performance of a series of CEs for the Pt(111)–O system, drawing on a large DFT-computed database of configurational energies. We have compared cluster expansions of various sizes and examined their predictive capabilities for formation energies, average and differential adsorption energies, and ground states. Pair-wise cluster expansions poorly predict all properties investigated, but larger CEs fare better. It was found that adsorption energy errors are significantly larger than the formation energy errors (typically 100s of meV per O rather than up to 10 meV per site) but that sufficiently larger cluster expansions reduce these errors.

In comparing these cluster expansions, short pairwise interaction as well as the 1–1–3 linear triplet were found to be important for accurately predicting properties; however, the selection of other specific figures was less important.

Adsorption-induced surface strain contributes significantly to adsorbate–adsorbate interactions. Surface relaxations due to adsorbates put the surface under compressive strain. This compressive strain results in relatively long-ranged, repulsive interactions for adsorbates beyond 1NN separations but reduces the repulsive interaction of 1NN adsorbates due to short-ranged cooperative relaxations. From a thermodynamic perspective, the long-ranged repulsions increase the probability of finding 1NN adsorbates, ironically acting as an additional pseudoattraction.

As alluded to in the Introduction, adsorbate–adsorbate interactions influence both the distributions/orderings of adsorbates at a surface as well as their adsorption energies. We have already demonstrated the ability of CEs to quickly calculate single oxygen, averaged, and differential adsorption energies. Additionally, the equilibrium spatial distributions of adsorbates at surfaces may be obtained using CEs coupled with the Monte Carlo method.²¹ The ability to precisely predict these quantities provides the foundation required for probing the quantity and type of energetically available dissociative O₂ adsorption sites, an avenue we are exploring to develop a better rate model for oxygen adsorption on the Pt(111) surface, and to gain insight into the types of dissociative O₂ adsorption sites that are available and important for catalytic oxidations and oxygen reduction reactions.³⁶

■ ASSOCIATED CONTENT

S Supporting Information. The DFT formation energy database, the ECI of the 27 figure CE, and a figure illustrating many-bodied interactions. This information is available free of charge via the Internet at <http://pubs.acs.org/>.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: (574) 631-8754. Fax: (574) 631-8366. E-mail: wschneider@nd.edu

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENT

Valuable conversations with Dr. Chao Wu, Dr. Jean-Sabin McEwen, Dr. Zhengzheng Chen, and Jason Bray are gratefully acknowledged. Support for this work comes from the National Science Foundation under grant CBET-0731020 and CBET-0730841.

■ REFERENCES

- (1) Langmuir, I. *J. Am. Chem. Soc.* **1918**, *40*, 1361–1403.
- (2) Langmuir, I. *J. Am. Chem. Soc.* **1932**, *54*, 2798–2832.
- (3) Kreuzer, H.; Jun, Z.; Payne, S.; Nichtl-Pecher, W.; Hammer, L.; Müller, K. *Surf. Sci.* **1994**, *303*, 1–15.
- (4) Bligaard, T.; Norskov, J.; Dahl, S.; Matthiesen, J.; Christensen, C.; Sehested, J. *J. Catal.* **2004**, *224*, 206–217.
- (5) Iddir, H.; Fong, D.; Zapol, P.; Fuoss, P.; Curtiss, L.; Zhou, G.; Eastman, J. *Phys. Rev. B* **2007**, *76*, 241404.
- (6) Müller, D.; Öberg, H.; Näslund, L.; Anniyev, T.; Ogasawara, H.; Pettersson, L.; Nilsson, A. *J. Chem. Phys.* **2010**, *133*, 224701.
- (7) Shustorovich, E. *Surf. Sci. Rep.* **1986**, *6*, 1–63.

- (8) Shelef, M.; McCabe, R. *Catal. Today* **2000**, *62*, 35–50.
- (9) Gandhi, H.; Graham, G.; McCabe, R. *J. Catal.* **2003**, *216*, 433–442.
- (10) Epling, W.; Campbell, L.; Yezerets, A.; Currier, N.; Parks, J. *Catal. Rev. Sci. Eng.* **2004**, *46*, 163–245.
- (11) Getman, R.; Schneider, W. *J. Phys. Chem. C* **2007**, *111*, 389–397.
- (12) Getman, R.; Xu, Y.; Schneider, W. *J. Phys. Chem. C* **2008**, *112*, 9559–9572.
- (13) Markovic, N. M.; Schmidt, T. J.; Stamenkovic, V.; Ross, P. *Fuel Cells* **2001**, *1*, 105–116.
- (14) Steininger, H.; Lehwald, S.; Ibach, H. *Surf. Sci.* **1982**, *123*, 1–17.
- (15) Mortensen, K.; Klink, C.; Jensen, F.; Besenbacher, F.; Stensgaard, I. *Surf. Sci. Lett.* **1989**, *220*, L701–L708.
- (16) Hawkins, J.; Weaver, J.; Asthagiri, A. *Phys. Rev. B* **2009**, *79*, 125434.
- (17) Campbell, C.; Ertl, G.; Kuipers, H.; Segner, J. *Surf. Sci.* **1981**, *107*, 220–236.
- (18) Parker, D.; Bartram, M.; Koel, B. *Surf. Sci.* **1989**, *217*, 489–510.
- (19) Yeo, Y.; Vattuone, L.; King, D. *J. Chem. Phys.* **1997**, *106*, 392–401.
- (20) Han, B.; Van der Ven, A.; Ceder, G.; Hwang, B. *Phys. Rev. B* **2005**, *72*, 205409.
- (21) Tang, H.; Van der Ven, A.; Trout, B. *Phys. Rev. B* **2004**, *70*, 045420.
- (22) Miller, S.; Kitchin, J. *Mol. Simul.* **2009**, *35*, 920–927.
- (23) Devarajan, S.; Hinojosa, J., Jr.; Weaver, J. *Surf. Sci.* **2008**, *602*, 3116–3124.
- (24) Weaver, J.; Chen, J.; Gerrard, A. *Surf. Sci.* **2005**, *592*, 83–103.
- (25) Miller, S.; Kitchin, J. *Surf. Sci.* **2009**, *603*, 794–801.
- (26) Van de Walle, A.; Ceder, G. *J. Phase Equilib.* **2002**, *23*, 348–359.
- (27) Van de Walle, A.; Asta, M.; Ceder, G. *Calphad* **2002**, *26*, 539–553.
- (28) Lerch, D.; Wieckhorst, O.; Hart, G.; Forcade, R.; Müller, S. *Modell. Simul. Mater. Sci. Eng.* **2009**, *17*, 055003.
- (29) Lerch, D.; Wieckhorst, O.; Hammer, L.; Heinz, K.; Müller, S. *Phys. Rev. B* **2008**, *78*, 121405.
- (30) Stampfl, C.; Kreuzer, H.; Payne, S.; Pfnür, H.; Scheffler, M. *Phys. Rev. Lett.* **1999**, *83*, 2993–2996.
- (31) McEwen, J.; Payne, S.; Stampfl, C. *Chem. Phys. Lett.* **2002**, *361*, 317–320.
- (32) Lazo, C.; Keil, F. *Phys. Rev. B* **2009**, *79*, 245418.
- (33) Chen, W.; Schmidt, D.; Schneider, W.; Wolverton, C. *J. Phys. Chem. C* **2011**, *115*, 17915–17924.
- (34) Chen, W.; Wolverton, C.; Schmidt, D.; Schneider, W. *Phys. Rev. B* **2011**, *83*, 075415.
- (35) Sanchez, J.; Ducastelle, F.; Gratias, D. *Phys. A* **1984**, *128*, 334–350.
- (36) Wu, C.; Schmidt, D.; Wolverton, C.; Schneider, W. *F. J. Catal.* **2011** in press.
- (37) Kresse, G.; Furthmüller, J. *Comput. Mater. Sci.* **1996**, *6*, 15–50.
- (38) Kresse, G.; Furthmüller, J. *Phys. Rev. B* **1996**, *54*, 1169.
- (39) Kresse, G.; Joubert, D. *Phys. Rev. B* **1999**, *59*, 1759.
- (40) Kresse, G.; Furthmüller, J. *Vienna ab-initio simulation package (VASP): The guide*; Institut für Materialphysik: Universität Wien, Vienna, 2007.
- (41) Perdew, J.; Wang, Y. *Phys. Rev. B* **1992**, *45*, 13244.
- (42) Blöchl, P. *Phys. Rev. B* **1994**, *50*, 17953.
- (43) Davey, W. *Phys. Rev.* **1925**, *25*, 753–761.
- (44) Murnaghan, F. *Proc. Natl. Acad. Sci. U. S. A.* **1944**, *30*, 244–247.
- (45) Birch, F. *Phys. Rev.* **1947**, *71*, 809–824.
- (46) Blöchl, P.; Jepsen, O.; Andersen, O. *Phys. Rev. B* **1994**, *49*, 16223.
- (47) Sette, F.; Stöhr, J.; Hitchcock, A. *J. Chem. Phys.* **1984**, *81*, 4906–4914.
- (48) Weber, A.; McGinnis, E. *J. Mol. Spectrosc.* **1960**, *4*, 195–200.
- (49) Linstrom, P.; Mallard, W. In *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*; Linstrom, P., Mallard, W., Eds.; National Institute of Standards and Technology: Gaithersburg, MD, 2010.
- (50) Stoicheff, B. *Can. J. Phys.* **1957**, *35*, 730–741.
- (51) van de Walle, A. *Alloy Theoretic Automated Toolkit (ATAT)*, 2.66 ed; Cal Tech: Pasadena, CA, 2008.
- (52) Bradshaw, A.; Richardson, N. *Pure Appl. Chem.* **1996**, *68*, 457–468.
- (53) Kreuzer, H.; Payne, S.; Drozdowski, A.; Menzel, D. *J. Chem. Phys.* **1999**, *110*, 6982–6999.
- (54) Schick, M.; Walker, J.; Wortis, M. *Phys. Rev. B* **1977**, *16*, 2205–2219.
- (55) Gsell, M.; Jakob, P.; Menzel, D. *Science* **1998**, *280*, 717–720.
- (56) Mavrikakis, M.; Hammer, B.; Nørskov, J. *Phys. Rev. Lett.* **1998**, *81*, 2819–2822.
- (57) Kitchin, J.; Nørskov, J.; Barteau, M.; Chen, J. *J. Chem. Phys.* **2004**, *120*, 10240–10246.
- (58) Kitchin, J.; Nørskov, J.; Barteau, M.; Chen, J. *Phys. Rev. Lett.* **2004**, *93*, 156801.
- (59) Ibach, H. *J. Vac. Sci. Technol., A* **1994**, *12*, 2240–2245.
- (60) Lin, X.; Ramer, N.; Rappe, A.; Hass, K.; Schneider, W.; Trout, B. *J. Phys. Chem. B* **2001**, *105*, 7739–7747.
- (61) Muller, S.; Wolverton, C.; Wang, L.; Zunger, A. *Phys. Rev. B* **1999**, *60*, 16448.
- (62) Ozolins, V.; Wolverton, C.; Zunger, A. *Phys. Rev. B* **1998**, *57*, 4816.
- (63) Ozolins, V.; Wolverton, C.; Zunger, A. *Phys. Rev. B* **1998**, *57*, 6427.
- (64) Wolverton, C.; Zunger, A. *Phys. Rev. Lett.* **1995**, *75*, 3162.
- (65) Laks, S.; Ferreira, L.; Froyen, S.; Zunger, A. *Phys. Rev. B* **1992**, *46*, 12587.

Anion Binding by Electron-Deficient Arenes Based on Complementary Geometry and Charge Distribution

Dong Young Kim, Inacrist Geronimo, N. Jiten Singh, Han Myoung Lee, and Kwang S. Kim*

Center for Superfunctional Materials, Department of Chemistry, Pohang University of Science and Technology, San 31, Hyojadong, Namgu, Pohang 790-784, Korea

S Supporting Information

ABSTRACT: Extended electron-deficient arenes are investigated as potential neutral receptors for poly-anions. Anion binds via σ interaction with extended arenes, which are composed solely of C and N ring atoms and CN substituents. As a result, the positive charge on the aromatic C is enhanced, consequently maximizing binding strength. Selectivity is achieved because different charge distributions can be obtained for target anions of a particular geometry. The halides F^- and Cl^- form the most stable complex with **6**, while the linear N_3^- interacts most favorably with **7**. The trigonal NO_3^- and tetrahedral ClO_4^- fit the 3-fold rotational axis of **6** but do not form stable complexes with **5** and **7**. The Y-shaped $HCOO^-$ forms complexes with **4**, **5**, and **7**, with the latter being the most stable. Thus, the anion complexes exhibit strong binding and the best geometrical fit between guest and host, reminiscent of Lego blocks.

INTRODUCTION

Anion recognition chemistry has grown as an important research area since the early 1960s because of the central role of anions in biological and chemical systems and its involvement in environmental pollution. Sensors commonly utilize ionic $(C-H)^+ \cdots X^-$ and neutral $(N-H) \cdots X^-$ hydrogen-bonding interactions in the detection of anions.^{1,2} On the other hand, the design of neutral anion receptors using Lewis acidic aromatic rings is a relatively recent research field.

Theoretical calculations of anion interaction with arenes in the gas phase show three types: (1) σ interaction, where the anion attacks a partially positive aromatic carbon, in effect changing the hybridization of the latter to sp^3 ; (2) anion- π interaction, primarily involving electrostatic (between the negative charge of the anion and the positive quadrupole moment of the arene) and dispersion effects; and (3) hydrogen-bond interaction arising from the increased acidity of the C-H donor due to electron-withdrawing groups.³⁻⁶ There is no established delineation between anion- π and weak covalent σ interaction; however, on the basis of extensive studies of mostly halide complexes, Hay et al.^{4b} proposed that the maximum density in the region between anion and arene, ρ_{max} is $<0.012 \text{ e } \text{\AA}^{-3}$ for anion- π , $0.012 \leq \rho_{max} \leq 0.100 \text{ e } \text{\AA}^{-3}$ for weak σ , and $\rho_{max} > 0.100 \text{ e } \text{\AA}^{-3}$ for strong σ interaction. It has been demonstrated that either the σ or the H bond complex is the global minimum conformation for the interaction between anion and Lewis acidic aromatic rings, particularly for F^- complexes. Nevertheless, experimental evidence of anion- π interaction has been reported, the stability of which can be attributed to environmental factors such as solvation and crystal packing effects.^{5c,7} Anion- π interaction has also been found in existing X-ray crystal structures, as discussed in a recent review.⁸

Receptors for polyatomic anions commonly utilize H bond interactions.⁹ It was demonstrated that H bond directionality can be exploited to achieve steric constraints for anion shape

recognition in urea-based host structures.¹⁰ On the other hand, anion recognition based on σ interaction has not been explored as extensively as other modes of interaction. Some experimental results show that this is a promising area of research for the development of an anion receptor. A neutral tripodal receptor composed of dinitroarenes has been reported to bind halides in solution by forming weak σ interactions.¹¹ Anion- σ binding by trinitrobenzene has been subsequently confirmed by an IR spectroscopic assay in the gas phase.¹² Fluoride was also shown to bond covalently to hexafluorobenzene through mass spectrometry.¹³ An attractive feature of the anion- σ interaction is that the resulting charge transfer absorption bands occur in the visible region, enabling colorimetric detection of anions.¹⁴ Possible weak σ interactions have been noted in some cases, as in polyazapyridinophane and hexasubstituted benzene-based receptors for NO_3^- , but have not been confirmed.^{9c,d}

Weak σ interaction, and possibly anion- π interaction, are generally enhanced with increased electron deficiency in the arene,³⁻⁶ and, as such, extended Lewis acidic arenes are viable receptors to strongly bind larger anions. Moreover, electron-deficient aromatic moieties, such as *N*-heteroaromatic rings, are characterized by nonuniform charge distribution, which can be exploited to specifically target anions of a particular charge distribution and geometry. In the present study, seven single and extended aromatic systems **1-7** are considered, composed solely of C and N, to find suitable anion receptors for halides (F^- and Cl^-) and linear (N_3^-), trigonal (NO_3^- , $HCOO^-$), and tetrahedral (ClO_4^-) polyatomic anions (Figure 1). In particular, strong binding was observed for F^- , Cl^- , NO_3^- , and ClO_4^- with **6**, $HCOO^-$ with **4** and **7**, and N_3^- with **7**.

Received: September 20, 2011

Published: November 23, 2011

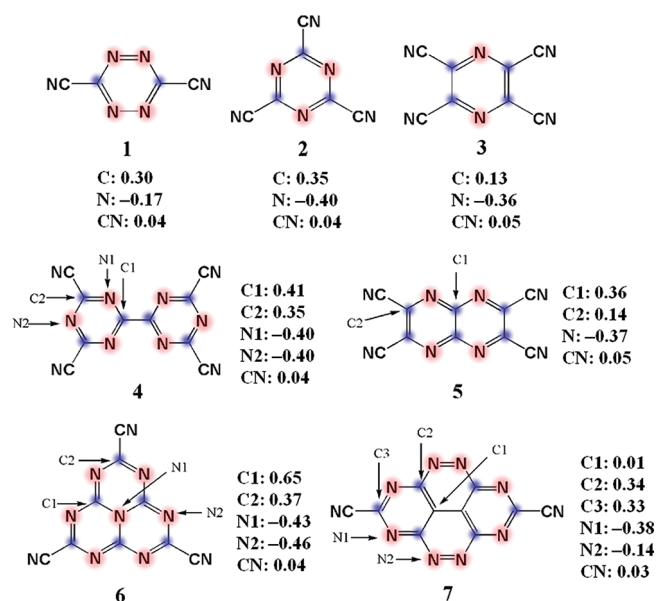


Figure 1. Seven types of arenes have different charge distributions due to the difference in electronegativities between the carbon and nitrogen atoms involved. Natural bond order (NBO) atomic charges of each system are shown in au. The net charge on the CN substituent is 0 au with positively charged C and negatively charged N ($|q(\text{C}/\text{N})| = 0.2\text{--}0.3$ au).

CALCULATION METHODS

Ab initio calculations were performed using Gaussian 03.¹⁵ The initial structures were optimized using Møller–Plesset second-order perturbation theory (MP2) with the 6-31+G* basis set and subsequently with the aug-cc-pVDZ (abbreviated as aVDZ) basis set. Frequency calculations were done at the MP2/aVDZ level to confirm minimum energy structures for all anion complexes of 1–5 and halide complexes of 6 and 7. Because of the computational cost, frequency calculations for 6–NO₃⁻, 6–ClO₄⁻, 7–HCOO⁻, and 7–N₃⁻ were performed at the MP2/6-31+G* level. The low-lying energy structures were corrected for basis set superposition error (BSSE) using the counterpoise (CP) method of Boys and Bernardi.¹⁶ The MP2/aug-cc-pVTZ (abbreviated as aVTZ) energy was determined to obtain the binding energies at the complete basis set (CBS) limit, which is based on the extrapolation method exploiting the fact that the electron correlation energy is proportional to N^{-3} for the aug-cc-pVTZ basis set.¹⁷ Natural population analysis (NPA) charges, where the charge distribution is derived from the basis functions representing the wave function,¹⁸ were calculated at the MP2/aVDZ level.

RESULTS AND DISCUSSION

A search on the Cambridge Structural Database (CSD, version 5.32 November 2010) revealed that 65 complexes containing arene moieties 1–3 exhibit possible σ interaction with halides at distances $R(\text{X}-\text{C}) = 3.3\text{--}3.4$ Å. In comparison, H bond interactions were found in 52 complexes at $R(\text{X}\cdots\text{H}) = 2.7\text{--}2.8$ Å and halide– π interactions in 45 complexes at vertical distance $R_v = 3.4\text{--}3.5$ Å. For interactions between arene moieties and NO₃⁻, 31 complexes in anion– π interactions (and possibly weak σ interactions) were found where the vertical distance between anion centroid and arene plane is $R_v = 3.3\text{--}3.6$ Å and

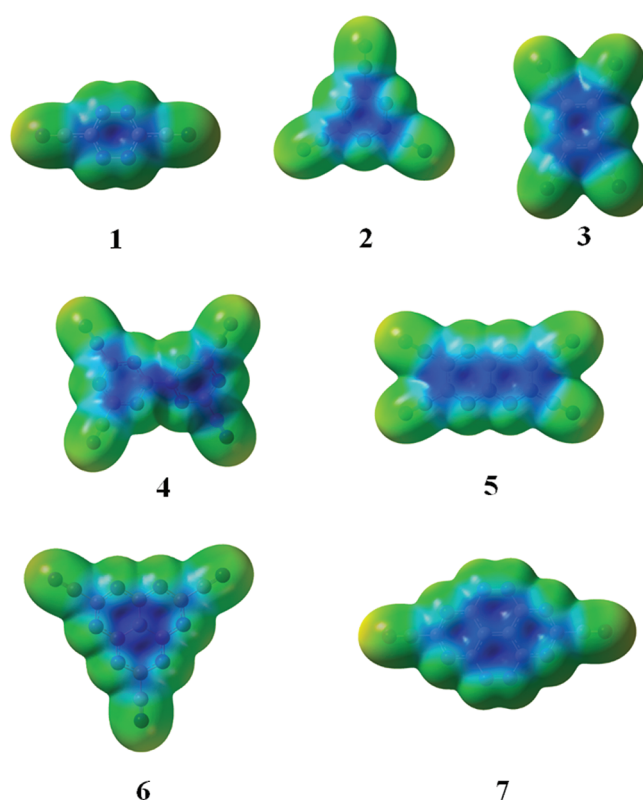


Figure 2. Molecular electrostatic potential (MEP) maps of 1–7, with the blue regions as positive and red as negative. Density rendered at 0.0020 ($\text{e} \text{Å}^{-3}$)^{1/2} isovalues.

the relative orientations of the molecular planes are distributed over 0–90°. In the case of ClO₄⁻, 70 complexes in anion– π interactions (and possibly weak σ interactions) were found where the vertical distance between anion centroid and arene plane is $R_v = 3.7\text{--}3.8$ Å, while most complexes adopt a T-shaped orientation (Supporting Information).

Figure 1 shows the natural bond order (NBO) charge distribution of seven single and extended aromatic systems 1–7, composed solely of C and N. An electron-withdrawing CN substituent (charge $q = 0$ au) instead of H ($q = 0.2$ au) makes the aromatic core more positive, resulting in enhanced anion binding strength. 1, 2, and 3 are single arenes whose aromatic cores are composed of four, three, and two negatively charged N atoms. The atomic charge of the aromatic C of 2 is the most positive ($q = 0.35$ au). 4–7 are extended arenes. In the cases of 4 and 5, the central aromatic carbons (C1) are highly positive, with $q = 0.41$ and 0.36 au, respectively. 6 has a highly positive carbon C1 ($q = 0.65$ au) surrounded by three N atoms. The central C1's of 7 are surrounded by four positively charged C2's ($q = 0.34$ au). 1–6 are synthesized; 7 is a hypothetical system to bind Y-shaped and linear anions. Figure 2 shows molecular electrostatic potential (MEP) maps of 1–7.

Optimized geometries for the anion complexes of single ring and extended systems are shown in Figure 3, with geometric and energetic parameters summarized in Table 1. Data for the other isomers can be found in the Supporting Information. The anion complexes are all true minima. The most stable F⁻ complexes with 1–3 involve strong covalent σ interaction between the aromatic C and F⁻, as evidenced by the significant charge transfer ($q_{\text{CT}} \approx 0.5$) (Table 1) and mixing of anion and arene

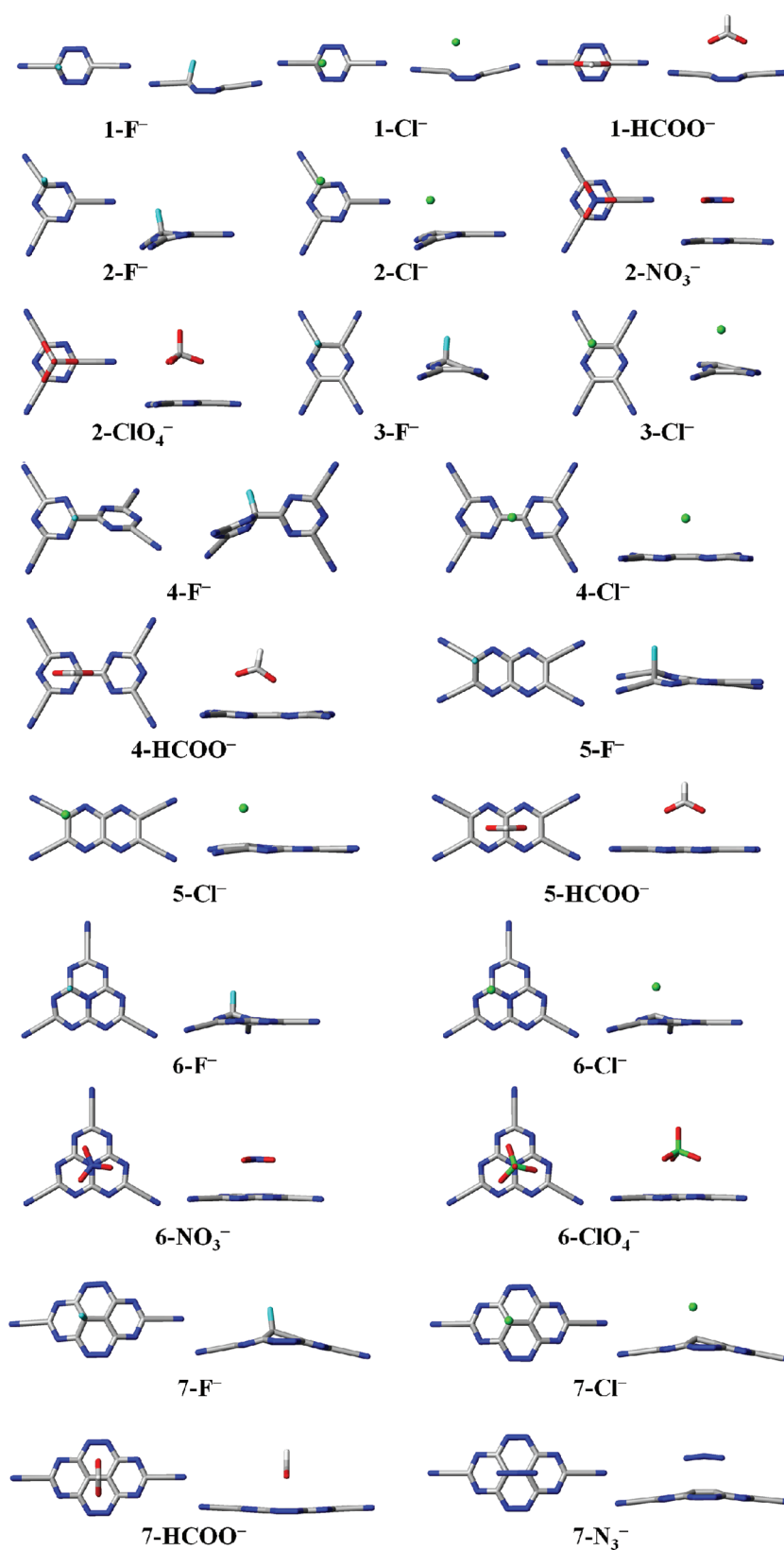


Figure 3. Optimized geometries (top and side views) of anion complexes of 1–7 at the BSSE-corrected MP2/aug-cc-pVDZ level.

orbitals. Bond distances are $\sim 1.5 \text{ \AA}$, and the aromatic C assumes a tetrahedral geometry. 1–F[−] has the largest binding energy (70.0 kcal/mol), in which the arene has the most positive

N atoms. Covalent σ complexes of 1–3 with Cl[−] have lower binding energies (35–40 kcal/mol); however, q_{CT} and mixing of orbitals are comparable to those of the F[−] complexes as shown

in Table 1 and Figure 4. Cl–C distances are longer (2.2–2.4 Å) than the experimental Cl–C(sp³) bond length (1.76 Å),¹⁹

Table 1. BSSE-Corrected MP2/CBS Binding Energies E_{MP2} (kcal/mol), Bond Distance R (Å), and Charge Transfer q_{CT} for the Global Minimum Structures of Anion Complexes of 1–7 in the Gas Phase

complex ^a	R , Å ^b	$-q_{\text{CT}}$ ^c	E_{MP2} , kcal/mol		
			aVDZ	aVTZ	CBS ^d
1–F [−]	1.49 (C)	0.52	63.25	68.01	70.01
1–Cl [−]	2.24 (C)	0.52	36.16	39.31	40.64
1–HCOO [−]	2.31 (C)	0.20	38.71	40.48	41.22
2–F [−]	1.49 (C)	0.52	63.39	67.89	69.78
2–Cl [−]	2.27 (C)	0.58	32.38	35.10	36.24
2–NO ₃ [−]	2.75 (C)	0.07	28.40	29.67	30.20
2–ClO ₄ [−]	2.80 (C)	0.05	25.89	27.06	27.56
3–F [−]	1.52 (C)	0.52	62.02	66.34	68.16
3–Cl [−]	2.42 (C)	0.40	36.02	38.29	39.24
4–F [−]	1.47 (C1)	0.52	66.34	71.04	73.01
4–Cl [−]	2.60 (C1)	0.30	41.18	43.71	44.77
4–HCOO [−]	2.63 (r), 2.32 (C1)	0.14	45.36	47.23	48.02
5–F [−]	1.52 (C2)	0.52	65.37	69.68	71.49
5–Cl [−]	2.51 (C2)	0.34	39.98	42.32	43.30
5–HCOO [−]	2.54 (r)	0.05	44.30	45.85	46.50
6–F [−]	1.45 (C1)	0.53	79.31	84.09	86.11
6–Cl [−]	2.00 (C1)	0.75	44.81	48.57	50.15
6–NO ₃ [−]	(r)	0.06	39.90	41.37	41.99
6–ClO ₄ [−]	2.71 (r)	0.04	37.66	39.26	39.93
7–F [−]	1.50 (C1)	0.50	69.29	73.14	74.76
7–Cl [−]	2.17 (C1)	0.59	41.87	44.41	45.49
7–HCOO [−]	(r)	0.06	46.72	48.45	49.18
7–N ₃ [−]	2.37 (C1)	0.30	49.73	51.85	52.74

^aFor other isomers, see Table S2 in the Supporting Information.

^bDistance between the aromatic carbon (Cx) or ring center (r) and the interacting atom in the anion. ^cCharge derived from Natural Population Analysis (NPA) and obtained by subtracting the calculated halide charge from the unit charge of the free halide. ^dValues in bold are the complexes with the highest binding energies for each anion.

although deformation is still observed in the aromatic ring. For comparison, F[−] and Cl[−] σ complexes of tetracyanobenzene (interaction with C–H instead of C–CN) have lower binding energies of 53.1 and 29.8 kcal/mol, respectively, while those of tricyanobenzene have binding energies of 44.1 and 22.7 kcal/mol, respectively, at the MP2/aVDZ level.^{4a}

The Y-shaped HCOO[−] also forms a σ complex with 1 (representative molecular orbital (MO) shown in Figure 4), with a slightly higher binding energy than 1–Cl[−] and an aromatic C–O bond length of 2.31 Å. On the other hand, 2 complexes with NO₃[−] and ClO₄[−] have much lower q_{CT} as compared to the corresponding halides. However, mixing of orbitals is observed in 2–NO₃[−] as in weak σ complexes. Because the distinction between anion– π and weak σ interactions is not well-defined for nonspherical anions, the type of interaction in 2–NO₃[−] and 2–ClO₄[−] cannot be identified with certainty. Binding energies for 2–NO₃[−] and 2–ClO₄[−] are 30.2 and 27.6 kcal/mol, respectively, lower than 2–Cl[−]. Anion– π complexes of Cl[−] and NO₃[−] with triazine, on the other hand, have much lower binding energies of \sim 7 kcal/mol, respectively, at the MP2/aVDZ level.^{3a} The H-bond complex of Cl[−]–triazine also has a lower binding energy of \sim 10 kcal/mol.^{4a}

Halides preferentially attack the C1 carbon of 4, which has a more positive charge. F[−] forms a strong covalent σ interaction with one of the C1 carbons (1.47 Å) with a binding energy of 73.0 kcal/mol. Cl[−], on the other hand, interacts with both C1 carbons (2.60 Å) with a binding energy of 44.8 kcal/mol. A potential energy scan along the C1–C1 axis in Figure 5 shows that the complex is most stable and charge transfer is most effective when Cl[−] is above the midpoint of the C1–C1 bond. 4–HCOO[−] has one of its O atoms above the center of the ring (2.63 Å) and the other interacting with one of the C1's (2.32 Å) and has a higher binding energy than the Cl[−] complex of 48.0 kcal/mol. The corresponding complex with N₃[−] is not a minimum structure (Table S2). In the fused, two-ring system 5, a C2 attack by

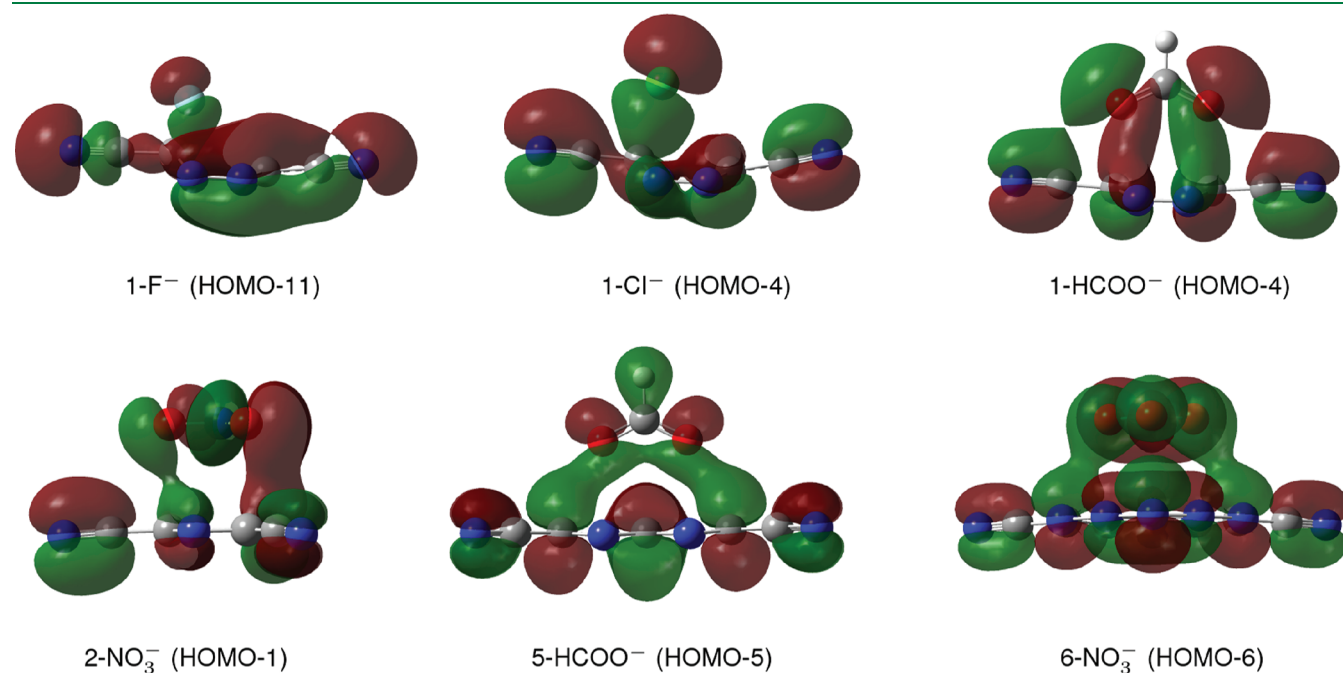


Figure 4. Representative molecular orbitals (MOs) of various anion complexes showing mixing of anion and arene orbitals. MOs rendered at 0.020 ($e \text{ \AA}^{-3}$)^{1/2} isovalues.

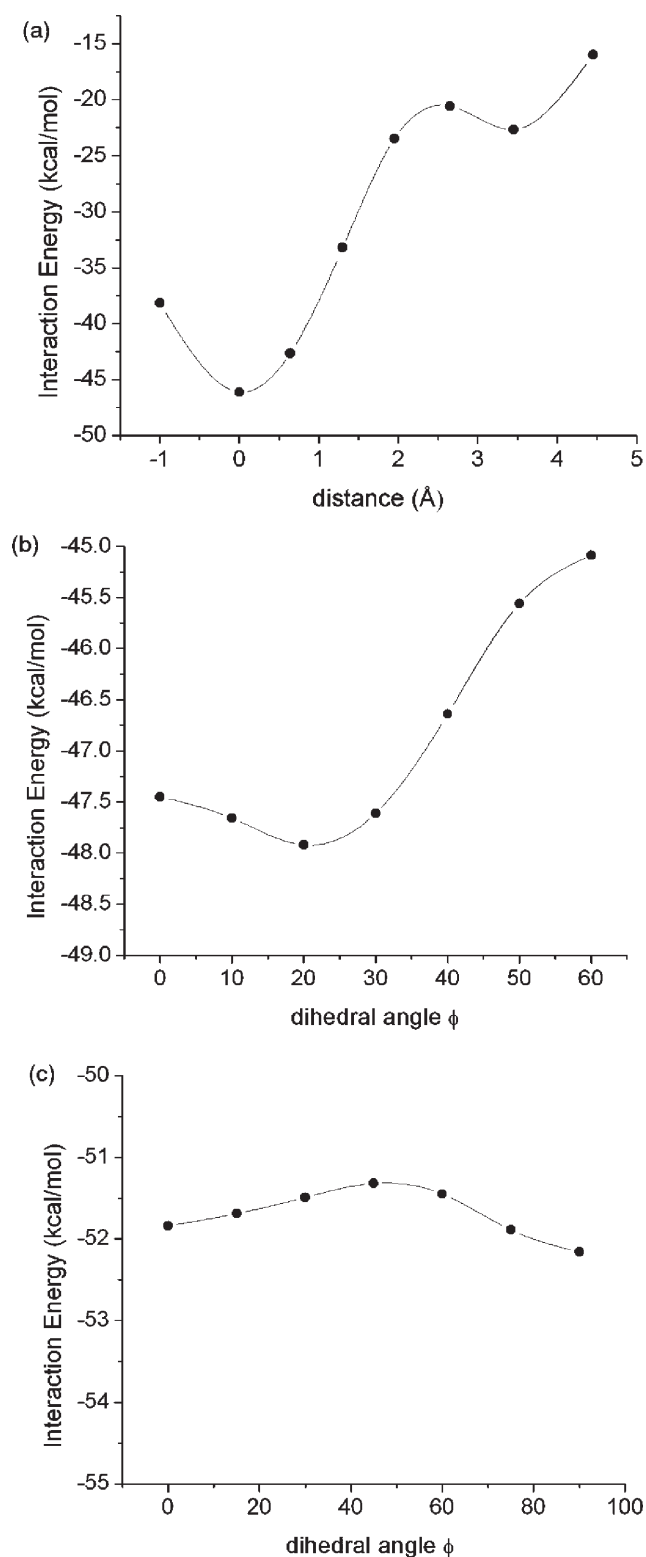


Figure 5. Potential energy scans for (a) 4-Cl⁻ along the C1-C1 axis (distance relative to the midpoint of the C1-C1 bond), (b) 6-NO₃⁻ along the O-N-N-C2 dihedral angle, and (c) 7-HCOO⁻ along the O-O-C1-C1 dihedral angle. Energies were calculated at the MP2/aug-cc-pVDZ level.

F⁻ and Cl⁻ is more favorable despite less positive charge because C1 is flanked by two negatively charged N atoms. The O atoms of

HCOO⁻ point toward the center of the ring at a distance of 2.54 Å. As with the complex 4, 5-HCOO⁻ has a larger binding energy of 46.5 kcal/mol than the Cl⁻ complexes. The representative MO of 5-HCOO⁻ shows mixing of orbitals despite the relatively low q_{CT} value. Complexes of 5 with NO₃⁻ and ClO₄⁻ have imaginary frequencies (Table S2).

NO₃⁻ and ClO₄⁻ form complexes with the three-ring system 6, with the O atoms interacting with the ring centers instead of the 3 C1s along the 3-fold rotational axis of 6. A potential energy scan along the O-N-N-C2 dihedral angle φ in the 6-NO₃⁻ complex (Figure 5) shows that an eclipsed conformation ($\varphi = 60^\circ$) results in a less stable complex. This is consistent with the molecular electrostatic potential (MEP) of 6 indicating that the electrostatic potential is most positive at the ring centers, despite the negative charge of the central nitrogen (-0.43 au). It has been demonstrated that the influence of substituents on the MEP is transmitted through-space, and hence does not necessarily imply local changes in the electron density.²⁰ Both F⁻ and Cl⁻ form strong σ interactions with C1. Binding energies with 6 are the highest among all of the F⁻ (86.1 kcal/mol), Cl⁻ (50.2 kcal/mol), NO₃⁻ (42.0 kcal/mol), and ClO₄⁻ (39.9 kcal/mol) complexes considered in the study. Corresponding complexes of Cl⁻, NO₃⁻, and ClO₄⁻ with cyameluric acid (=O substituent instead of CN) have been previously reported, but these have much lower binding energies.^{3b}

7-F⁻ and 7-Cl⁻ involve interaction with C1 with binding energies of 74.8 and 45.5 kcal/mol, respectively. The linear anion N₃⁻ has a strong σ interaction with the C1 atoms with the binding energy of 52.7 kcal/mol. In contrast, the O atoms of HCOO⁻ interact with the ring centers, as in 5-HCOO⁻, with a binding energy of 49.2 kcal/mol. However, a potential energy scan along the O-O-C1-C1 dihedral angle of the 7-HCOO⁻ complex shows that the energy difference between different orientations, including one for which the O atoms point toward the C1 atoms ($\varphi = 0^\circ$), is not significant (less than 1 kcal/mol). As shown in the MEPs of 5 and 7, the region above the ring centers also has the most positive electrostatic potential. A 7-NO₃⁻ complex with the same geometry as 7-HCOO⁻ is also a minimum albeit with a lower binding energy. On the other hand, the isomers of 7-HCOO⁻ and 7-NO₃⁻ wherein the O atoms interact with C1 have imaginary frequencies. The corresponding anion complex of 7 with the tetrahedral ClO₄⁻ is not a minimum structure as well (Table S2).

CONCLUSIONS

In summary, strong binding of anions is exhibited by extended Lewis acidic aromatic rings, and selective recognition is accomplished by matching the geometry and charge distribution between anion and arene. The halides F⁻ and Cl⁻ form the most stable complex with 6, while the linear N₃⁻ only forms a minimum structure with 7. These complexes are characterized by σ interaction with the most positive aromatic C. The Y-shaped HCOO⁻ forms the most stable complex with 7, where the O atoms interact with the ring centers. The trigonal NO₃⁻ and tetrahedral ClO₄⁻ fit the 3-fold rotational axis of 6. However, complexes of these anions with 5 and 7 are not minimum structures. The type of interaction for HCOO⁻, NO₃⁻, and ClO₄⁻ cannot be identified conclusively because characteristics of anion- π and weak σ interaction are not well-defined for nonspherical anions. However, they are characterized by weak charge transfer (<0.1 au), so they are less susceptible to solvent

influences.^{5c} The effect of solvation on the stability of σ complexes is beyond the scope of the present study; however, the role of explicit solvation on anion complexation has been reported in a previous study.^{5c} It can be seen that the most stable complexes show the best geometrical fit between guest and host, reminiscent of Lego blocks. Thus, tailoring extended electron-deficient arenes to bind polyanions of a specific geometry and charge distribution provides a strategic method in the design of selective neutral anion receptors and anion-based self-assembly architectures.

■ ASSOCIATED CONTENT

Supporting Information. Structures and interaction energies of low-lying energy complexes. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: kim@postech.ac.kr

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENT

This work was supported by KRF (National Honor Scientist Program, 2010-0020414, WCU: R32-2008-000-10180-0) and KISTI (KSC-2011-G3-02).

■ REFERENCES

- (1) (a) Schulze, B.; Friebe, C.; Hager, M. D.; Günther, W.; Köhn, U.; Jahn, B. O.; Görls, H.; Schubert, U. S. *Org. Lett.* **2010**, *12*, 2710. (b) Yoon, J.; Kim, K. S.; Singh, N. J.; Kim, K. S. *Chem. Soc. Rev.* **2006**, *35*, 355. (c) Chellapan, K.; Singh, N. J.; Hwang, I.-C.; Lee, J. W.; Kim, K. S. *Angew. Chem.* **2005**, *117*, 2959. (d) Ihm, H.; Yun, S.; Kim, H. G.; Kim, J. K.; Kim, K. S. *Org. Lett.* **2002**, *4*, 2897.
- (2) (a) Nielsen, K. A.; Jeppesen, J. O.; Levillain, E.; Becher, J. *Angew. Chem., Int. Ed.* **2003**, *42*, 187. (b) Cho, E. J.; Moon, J. W.; Ko, S. W.; Lee, J. Y.; Kim, S. K.; Yoon, J.; Nam, K. C. *J. Am. Chem. Soc.* **2003**, *125*, 12376. (c) Jose, D. A.; Kumar, D. K.; Ganguly, B.; Das, A. *Org. Lett.* **2004**, *6*, 3445. (d) Mascal, M.; Yakovlev, I.; Nikitin, E. B.; Fetting, J. C. *Angew. Chem., Int. Ed.* **2006**, *45*, 4628. (e) Sessler, J. L.; Cai, J.; Gong, H.-Y.; Yang, X.; Arambula, J. F.; Hay, B. P. *J. Am. Chem. Soc.* **2010**, *132*, 14058.
- (3) (a) Mascal, M.; Armstrong, A.; Bartberger, M. D. *J. Am. Chem. Soc.* **2002**, *124*, 6274. (b) Alkorta, I.; Rozas, I.; Elguero, J. *J. Am. Chem. Soc.* **2002**, *124*, 8593. (c) Clements, A.; Lewis, M. *J. Phys. Chem. A* **2006**, *110*, 12705. (d) Schottel, B. L.; Chifotides, H. T.; Dunbar, K. R. *Chem. Soc. Rev.* **2008**, *37*, 68. (e) Ran, J.; Hobza, P. *J. Chem. Theory Comput.* **2009**, *5*, 1180. (f) Wheeler, S. E.; Houk, K. N. *J. Phys. Chem. A* **2010**, *114*, 8658.
- (4) (a) Berryman, O. B.; Bryantsev, V. S.; Stay, D. P.; Johnson, D. W.; Hay, B. P. *J. Am. Chem. Soc.* **2007**, *129*, 48. (b) Hay, B. P.; Bryantsev, V. S. *Chem. Commun.* **2008**, 2417. (c) Hay, B. P.; Custelcean, R. *Cryst. Growth Des.* **2009**, *9*, 2539.
- (5) (a) Kim, D.; Tarakeswar, P.; Kim, K. S. *J. Phys. Chem. A* **2004**, *108*, 1250. (b) Kim, D. Y.; Singh, N. J.; Kim, K. S. *J. Chem. Theory Comput.* **2008**, *4*, 1401. (c) Kim, D. Y.; Singh, N. J.; Lee, J. W.; Kim, K. S. *J. Chem. Theory Comput.* **2008**, *4*, 1162.
- (6) (a) Quiñonero, D.; Garau, C.; Rotger, C.; Frontera, A.; Ballester, P.; Costa, A.; Deyà, P. M. *Angew. Chem., Int. Ed.* **2002**, *41*, 3389. (b) Garau, C.; Frontera, A.; Quiñonero, D.; Ballester, P.; Costa, A.; Deyà, P. M. *J. Phys. Chem. A* **2004**, *108*, 9423. (c) Quiñonero, D.; Garau, C.; Frontera, A.; Ballester, P.; Costa, A.; Deyà, P. M. *J. Phys. Chem. A* **2005**, *109*, 4632. (d) Frontera, A.; Quiñonero, D.; Costa, A.; Ballester, P.; Deyà, P. M. *New J. Chem.* **2007**, *31*, 556.
- (7) (a) Demeshko, S.; Dechert, S.; Meyer, F. *J. Am. Chem. Soc.* **2004**, *126*, 4508. (b) de Hoog, P.; Gamez, P.; Mutikainen, I.; Turpeinen, U.; Reedijk, J. *Angew. Chem., Int. Ed.* **2004**, *43*, 5815. (c) Schottel, B. L.; Chifotides, H. T.; Shatruk, M.; Chouai, A.; Pérez, L. M.; Bacsa, J.; Dunbar, K. R. *J. Am. Chem. Soc.* **2006**, *128*, 5895. (d) Wang, D.-X.; Zheng, Q.-Y.; Wang, Q.-Q.; Wang, M.-X. *Angew. Chem., Int. Ed.* **2008**, *47*, 7595. (e) Barrios, L. A.; Aromí, G.; Frontera, A.; Quiñonero, D.; Deyà, P. M.; Gamez, P.; Roubeau, O.; Shotton, E. J.; Teat, S. *J. Inorg. Chem.* **2008**, *47*, 5873. (f) Gil-Ramirez, G.; Escudero-Adan, E. C.; Benet-Buchholz, J.; Ballester, P. *Angew. Chem., Int. Ed.* **2008**, *47*, 4114. (g) Perez-Velasco, A.; Gortea, V.; Matile, S. *Angew. Chem., Int. Ed.* **2008**, *47*, 921. (h) de Hoog, P.; Robertazzi, A.; Mutikainen, I.; Turpeinen, U.; Gamez, P.; Reedijk, J. *Eur. J. Inorg. Chem.* **2009**, 2684. (i) Gural'skiy, I. A.; Escudero, D.; Frontera, A.; Solntsev, P. V.; Rusanov, E. B.; Chernega, A. N.; Krautscheid, H.; Domasevitch, K. V. *Dalton Trans.* **2009**, 2856. (j) Albrecht, M.; Müller, M.; Mergel, O.; Rissanen, K.; Valkonen, A. *Chem.-Eur. J.* **2010**, *16*, 5062. (k) Xu, Z.; Singh, N. J.; Kim, S. K.; Spring, D. R.; Kim, K. S.; Yoon, J. *Chem.-Eur. J.* **2010**, *17*, 1163. (l) Dawson, R. E.; Hennig, A.; Weimann, D. P.; Emery, D.; Rauikumar, V.; Montenegro, J.; Takeuchi, T.; Gabutti, S.; Mayor, M.; Mareda, J.; Schalley, C. A.; Matile, S. *Nature Chem.* **2010**, *2*, 533. (m) Chifotides, H. T.; Schottel, B. L.; Dunbar, K. R. *Angew. Chem., Int. Ed.* **2010**, *49*, 7202. (n) Garcia-Raso, A.; Albertí, F. M.; Fiol, J. J.; Lagos, Y.; Torres, M.; Molins, E.; Mata, I.; Estarellas, C.; Frontera, A.; Quiñonero, D.; Deyà, P. M. *Eur. J. Org. Chem.* **2010**, 5171.
- (8) Frontera, A.; Gamez, P.; Mascal, M.; Mooibroek, T. J.; Reedijk, J. *Angew. Chem., Int. Ed.* **2011**, *50*, 9564.
- (9) (a) Bryantsev, V. S.; Hay, B. P. *J. Am. Chem. Soc.* **2006**, *128*, 2035. (b) Blondeau, P.; Segura, M.; Pérez-Fernández, R.; de Mendoza, J. *Chem. Soc. Rev.* **2007**, *36*, 198. (c) Valencia, L.; Bastida, R.; Garcia-España, E.; de Julián-Ortiz, J. V.; Llinares, J. M.; Macías, A.; Lourido, P. P. *Cryst. Growth Des.* **2010**, *10*, 3418. (d) Arunachalam, M.; Ghosh, P. *Org. Lett.* **2010**, *12*, 328.
- (10) (a) Hay, B. P.; Firman, T. K.; Moyer, B. A. *J. Am. Chem. Soc.* **2005**, *127*, 1810. (b) Hay, B. P.; Dixon, D. A.; Bryan, J. C.; Moyer, B. A. *J. Am. Chem. Soc.* **2002**, *124*, 182. (c) Amendola, V.; Esteban-Gómez, D.; Fabbri, L.; Licchelli, M. *Acc. Chem. Res.* **2006**, *39*, 343.
- (11) (a) Berryman, O. B.; Sather, A. C.; Hay, B. P.; Meisner, J. S.; Johnson, D. W. *J. Am. Chem. Soc.* **2008**, *130*, 10895. (b) Berryman, O. B.; Johnson, D. W. *Chem. Commun.* **2009**, 3143.
- (12) Chiavarino, B.; Crestoni, M. E.; Fornanini, S.; Lanucara, F.; Lemaire, J.; Maître, P.; Scuderi, D. *Chem.-Eur. J.* **2009**, *15*, 8185.
- (13) Hiraoka, K.; Mizuse, S.; Yamabe, S. *J. Phys. Chem.* **1987**, *91*, 5294.
- (14) Rosokha, Y. S.; Lindeman, S. V.; Rosokha, S. V.; Kochi, J. K. *Angew. Chem., Int. Ed.* **2004**, *43*, 4650.
- (15) 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.
- (16) (a) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553. (b) Simon, S.; Duran, M.; Dannenberg, J. J. *J. Chem. Phys.* **1996**, *105*, 11024.

- (17) (a) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *J. Chem. Phys.* **1997**, *106*, 9639. (b) Min, S. K.; Lee, E. C.; Lee, H. M.; Kim, D. Y.; Kim, D.; Kim, K. S. *J. Comput. Chem.* **2008**, *29*, 1208. (c) Lee, E. C.; Kim, D.; Jurecka, P.; Tarakeshwar, P.; Hobza, P.; Kim, K. S. *J. Phys. Chem. A* **2007**, *111*, 3446.
- (18) Reed, A. E.; Weinstock, R. B.; Weinhold, F. *J. Chem. Phys.* **1985**, *83*, 735.
- (19) Zhang, N.; Blowers, P.; Farrell, J. *Environ. Sci. Technol.* **2005**, *39*, 612.
- (20) Wheeler, S. E.; Houk, K. N. *J. Chem. Theory Comput.* **2009**, *5*, 2301.

Noncovalent Interactions in SIESTA Using the vdW-DF Functional: S22 Benchmark and Macrocyclic Structures

Damien J. Carter^{†,‡} and Andrew L. Rohl^{*,†,‡}

[†]Nanochemistry Research Institute, Department of Chemistry, Curtin University of Technology, GPO Box U1987, Perth, WA, Australia, 6845

[‡]iVEC, 26 Dick Perry Avenue, Technology Park, Kensington, WA, Australia 6151

ABSTRACT: We investigate the performance of the vdW-DF functional of Dion et al. implemented in the SIESTA code. In particular, the S22 data set and several calixarene-based host–guest structures are examined to assess the performance of the functional. The binding energy error statistics for the S22 data set reveal that the vdW-DF functional performs very well when compared to a range of other methods of treating dispersion in density functional theory, and to vdW-DF implementations in other codes. For the calixarene host–guest structures, the structural properties and binding energies are compared to previous experimental and computational studies, and in most cases we find that vdW-DF provides superior results to other computational studies.

INTRODUCTION

Ab initio quantum mechanical methods, in particular density functional theory (DFT), arguably offer the most accurate methods for determining the stability and properties of structures; however, they are limited by the size of systems that can be examined. A number of DFT codes, such as SIESTA,¹ use localized basis sets, pseudopotentials, and other features such as linear scaling that facilitate faster calculations, enabling investigations of much larger systems.

Up until a few years ago, these relatively fast DFT calculations were not being commonly used for soft matter or biomolecular or molecular crystals. This is because standard DFT methods lacked a description of van der Waals (vdW), or dispersion, forces, which can be large and important contributions in these types of systems. The “gold standard” of chemical accuracy in quantum methods is arguably the perturbatively corrected coupled cluster CCSD(T) method;² however, the computational effort scales by a formal cost (where N is the system size) of $O(N^7)$, which limits its applicability for these soft matter applications. As a result, in the past few years, there has been an explosion of methods that have been developed for DFT to provide ever increasingly more accurate descriptions of the dispersion forces and have been generally termed DFT-D^{3–5} methods.

A popular proposal for including dispersion forces has been to add an empirical correction using interatomic potentials of the form C_6R^{-6} ,^{3,6–9} with parameters derived from fitting to quantum mechanical calculations. These types of corrections are added directly to standard exchange correction functionals, and a “-D” is appended to the name, such as BLYP-D³ and B97-D⁴ or to hybrid functionals such as B3LYP-D.⁸ A number of methods have been published that include medium-range dispersion forces in conventional semilocal DFT, with hybrid meta-GGA methods such as X3LYP,¹⁰ ω B97X-D,¹¹ M06,¹² and PW6B95.¹³ Another popular option has been to account for dispersion by incorporating correlation components from wave function theory, sometimes called double hybrid density functionals, with methods such as B2PLYP¹⁴ and XYG3.¹⁵

Another approach has been to generate dispersion coefficients based on the exchange-hole dipole method (XDM). In the XDM^{16–18} approach, dispersion interactions are modeled by examining the instantaneous dipole that arises between an electron and its exchange hole. A more complex approach involves the development of explicitly nonlocal correlation functionals from first principles. Examples of this include the vdW-DF,¹⁹ vdW-DF2,²⁰ and VV09²¹ functionals. These methods are potentially more accurate than the parametrized methods mentioned above, particularly for vdW interactions that depend on their chemical environment.²² The vdW-DF functional has recently been implemented²² in the SIESTA code and successfully been used to examine binding energies in double-walled carbon nanotubes²² and to calculate the properties of metal organic framework (MOF) materials.^{23,24}

In this work, we performed calculations on the S22 test set of molecules developed by Jurecka et al.²⁵ for assessing vdW interactions to assess the accuracy of results of the vdW-DF implementation in the SIESTA code. In particular, we examined the effect of several basis sets and the effect of geometry relaxation and compare the binding energies to literature reports for the S22 test set using a variety of different density functional methods, including comparisons to results from vdW-DF implementations in other software codes. We also calculated the structures of two calixarene inclusion compounds, namely, *p*-tert-butylcalix[4]arene·CS₂ and *p*-tert-butylcalix[4]arene·toluene, and compare the results to previous theoretical and experimental studies.

METHODOLOGY

All DFT calculations were performed using the SIESTA¹ code. DFT-D calculations were performed using vdW-DF,¹⁹ as described by Roman-Perez and Soler.²² For comparison, standard DFT calculations were also performed using the PBE²⁶ functional.

Received: September 27, 2011

Published: November 16, 2011

Norm-conserving pseudopotentials of Troullier and Martins²⁷ were used with the valence electron configurations of hydrogen 1s, carbon 2s²2p², nitrogen 2s²2p³, oxygen 2s²2p⁴, and sulfur 3s²3p⁴. Hartree and exchange correlation energies were evaluated on a uniform real-space grid of points with a defined maximum kinetic energy of 300 Ry. For basis set generation, we used soft confinement potentials²⁸ to generate both double- ζ and triple- ζ plus polarization basis sets. Standard basis sets were generated, where the numerical atomic orbitals were radially confined to an extent that induces an energy shift in each orbital of 0.001 Ry (we refer to these as the DZ, DZP, or TZP basis for double- ζ , double- ζ polarized, and triple- ζ polarized basis sets). The other basis set was of triple- ζ polarized quality (herein referred to as TZP-L) and used an 8 Bohr cutoff for all orbital types (s, p, d, and f) and an explicit polarization orbital defined as a single- ζ of $(l + 1)$ angular momentum, also with an 8 Bohr cutoff.

We use the S22 set of complexes of common molecules to examine the accuracy of our vdW-DF calculations. The S22 set can be grouped into three subgroups based on their noncovalent interactions: (i) hydrogen-bonded complexes, (ii) complexes with predominant dispersion contributions, and (iii) mixed complexes in which electrostatic and dispersion contributions are similar in magnitude. The S22 reference geometries are taken from Jurecka et al.,²⁵ where all geometries were optimized at either the CCSD(T) or MP2 level (and in several cases where hydrogen positions were not reported for reference structures, they optimized the hydrogen atom positions at the DFT B3LYP level). The binding energies of the S22 set have recently been revised (the reference geometries are unchanged) by Takatani et al.,²⁹ who used a larger and more complete basis set than the original work, so we will compare our binding energies to both the original S22 binding energies of Jurecka et al.²⁵ and the new S22A binding energies of Takatani et al.²⁹ Our calculations of the S22 set are performed using both the reference geometries and with full geometry relaxations.

While many literature studies of test sets of compounds include an estimate for basis sets superposition errors (BSSE), usually via the counterpoise correction (CP) method, there are just as many literature studies that do not report any BSSE corrections with their binding energies. The counterpoise method can be problematic because, although it can improve the accuracy for very small basis sets (smaller than 6-31+G(d,p)), it can lead to less accurate results for moderate and large basis sets.³⁰ In more complex systems such as biopolymers, trimers, and other soft materials, the CP method can be impractical or ambiguous.³¹ Grimme and co-workers^{3,4,32} argue that with good quality basis sets, such as polarized triple- ζ basis sets, BSSE effects are small and are not required, and this is the approach that we will follow. We calculated a BSSE correction for several S22 examples and found it was of similar magnitude to that reported by Antony and Grimme,³² supporting our decision not to report BSSE calculations in this paper.

For calculations of the structures of *p*-tert-butylcalix[4]arene·CS₂ and *p*-tert-butylcalix[4]arene·toluene, we use the TZP-L basis set and the other computational parameters listed above. We optimized the structures in both the gas phase and solid state (fixed at the experimental lattice parameters and fully relaxed), using the *p*-tert-butylcalix[4]arene·CS₂ crystal structure of Schatz et al.³³ and the *p*-tert-butylcalix[4]arene·toluene crystal structure of Arduini et al.³⁴

RESULTS AND DISCUSSION

A. S22 Data Set. In Table 1, we report the binding energies for calculations of the S22 data set for a range of basis sets and the GGA functionals with (vdW-DF) and without (PBE) dispersion, at the reference and fully relaxed geometries. The reference energies in Table 1 are the S22A binding energies from Takatani et al.²⁹

There are some clear trends in the binding energies in Table 1 with, for example, PBE overestimating the binding energies for hydrogen bonded complexes. As a visual aid to clearly show the trends in Table 1, we have plotted the difference between the binding energies from our calculations (ΔE) and the binding energies of the reference S22A data set (ΔE_{ref}), as shown in Figure 1. For the PBE functional, we see a clear trend of overestimation of binding energies for hydrogen bonded complexes and an underestimation for complexes with predominantly dispersion interactions or for mixed interaction compounds. This is the case for both the reference geometries and fully relaxed structures. Due to some fortuitous error cancellation, PBE still performs reasonably well for some compounds, such as the ammonia dimer or the ethene–ethyne dimer. Jurecka et al.³⁵ also reported similar behavior in their calculations of the S22 data set. GGA functionals like PBE are known to typically overestimate the strength of hydrogen bonds,³⁶ and we found similar behavior in our previous SIESTA calculations of the strongly hydrogen bound potassium dihydrogen phosphate system.^{37,38}

For the vdW-DF functional using TZP and TZP-L basis sets, we find the opposite behavior of that of the PBE functional, in general underestimating the binding energy for hydrogen bonded complexes and overestimating for predominantly dispersion and mixed complexes, although the deviation is much less than for the PBE functional. Using the slightly smaller DZP basis sets, the vdW-DF generally overestimates (although some are still underbound) the binding of all complexes. Using the DZ basis set, the smallest basis set we examined, the results for the dispersion and mixed complexes are similar in magnitude and direction to the larger basis set results; however for the hydrogen bonded complexes, there is a large problem of overbinding. The fully relaxed geometries exhibit similar trends to those with the geometries fixed at the S22 reference geometry.

To quantify the trends in the binding energy results for the S22 data set as a whole, we compute three quantities—namely, the mean deviation (MD), mean absolute deviation (MAD), and root-mean-square deviation (RMSD). These are defined in the following way:

$$\text{MD} = \frac{1}{n} \sum_{\text{sys}}^n (\Delta E - \Delta E_{\text{ref}}),$$

$$\text{MAD} = \frac{1}{n} \sum_{\text{sys}}^n |\Delta E - \Delta E_{\text{ref}}|,$$

$$\text{RMSD} = \sqrt{\frac{1}{n} \sum_{\text{sys}}^n (\Delta E - \Delta E_{\text{ref}})^2}$$

In Table 2, we compare the MD, MAD, and RMSD for our calculated binding energies based on the S22A reference energies and compare the results to a recent study by Burns et al.,⁴⁰ who reported a comprehensive investigation of the S22 data set with a wide range of exchange-correlation (XC) functionals. In Table 2, we report only a selection of different XC functional results from

Table 1. Binding Energies (kcal/mol) for the S22 Data Set^a

functional	basis set	ref	reference geometry					relaxed geometry				
			PBE		vdW-DF			PBE		vdW-DF		
			TZP	DZ	DZP	TZP	TZP-L	TZP	DZ	DZP	TZP	TZP-L
hydrogen bonded complexes												
1	(NH ₃) ₂	-3.17	-3.46	-3.96	-3.10	-3.04	-3.01	-3.52	-3.91	-3.26	-3.14	-3.12
2	(H ₂ O) ₂	-5.02	-6.49	-7.25	-5.57	-5.49	-5.36	-6.86	-8.22	-5.95	-5.76	-5.56
3	formic acid dimer (C _{2h})	-18.61	-22.16	-23.04	-20.32	-18.53	-18.82	-22.46	-23.57	-18.58	-16.90	-17.68
4	formamide dimer (C _{2h})	-15.96	-17.44	-18.60	-16.68	-15.16	-15.37	-17.78	-19.83	-16.36	-14.75	-15.46
5	uracil dimer (C _{2h})	-20.65	-21.37	-24.17	-21.17	-19.64	-20.05	-21.22	-24.54	-19.99	-18.35	-19.32
6	2-pyridoxine-2-aminopyridine	-16.71	-18.57	-22.18	-17.85	-17.02	-17.36	-19.25	-23.56	-17.37	-16.47	-16.98
7	adenine-thymine WC	-16.37	-17.91	-23.61	-17.46	-16.42	-16.97	-18.28	-22.56	-17.09	-15.95	-16.58
complexes with predominant dispersion contributions												
8	(CH ₄) ₂ (D _{3d})	-0.53	-0.48	-1.33	-1.13	-1.22	-0.96	-0.44	-3.91	-1.13	-1.21	-0.96
9	(C ₂ H ₄) ₂ (D _{2d})	-1.51	-0.77	-1.43	-1.50	-1.65	-1.59	-0.84	-8.22	-1.67	-1.73	-1.76
10	benzene-CH ₄ (C ₃)	-1.50	-0.31	-1.26	-1.37	-1.44	-1.78	-0.76	-23.57	-1.78	-1.85	-2.04
11	benzene dimer (C _{2h})	-2.73	0.02	-3.24	-3.23	-4.01	-3.59	-0.63	-19.83	-3.60	-4.18	-4.00
12	pyrazine dimer (C _{2v})	-4.42	-1.43	-5.45	-4.86	-5.31	-5.18	-1.93	-24.54	-5.31	-5.61	-5.57
13	uracil dimer (C _s)	-10.12	-6.04	-9.55	-11.87	-11.97	-12.09	-5.63	-23.56	-11.18	-10.82	-11.37
14	indole benzene	-5.22	-0.47	-4.97	-5.07	-6.00	-5.61	-2.16	-22.56	-5.57	-6.29	-6.01
15	adenine thymine (stack)	-12.23	-5.75	-12.58	-12.81	-13.38	-13.40	-6.06	-3.91	-13.19	-13.42	-13.43
mixed complexes												
16	ethene-ethyne (C _{2v})	-1.53	-1.64	-1.33	-1.13	-1.94	-2.14	-1.80	-1.32	-1.68	-2.00	-2.21
17	benzene H ₂ O (C _s)	-3.28	-3.46	-1.43	-1.67	-4.24	-3.91	-3.72	-1.70	-4.50	-4.34	-3.95
18	benzene NH ₃ (C _s)	-2.35	-1.75	-1.26	-1.78	-2.69	-2.80	-1.86	-1.63	-2.65	-2.82	-2.88
19	benzene HCN (C _s)	-4.46	-3.62	-3.24	-3.60	-4.37	-4.76	-3.71	-3.68	-4.18	-4.44	-4.84
20	benzene dimer (C _{2v})	-2.74	-1.24	-5.45	-5.31	-3.19	-3.44	-1.36	-5.91	-2.76	-3.22	-4.00
21	indole benzene T-shape	-5.73	-3.84	-9.55	-11.18	-5.87	-6.47	-3.82	-10.40	-5.64	-5.88	-6.39
22	phenol dimer	-7.05	-6.14	-4.97	-5.57	-7.29	-7.91	-6.60	-5.60	-7.19	-7.45	-8.26

^aThe reference binding energies (Ref.) are the S22A binding energies from Takatani *et al.*²⁹

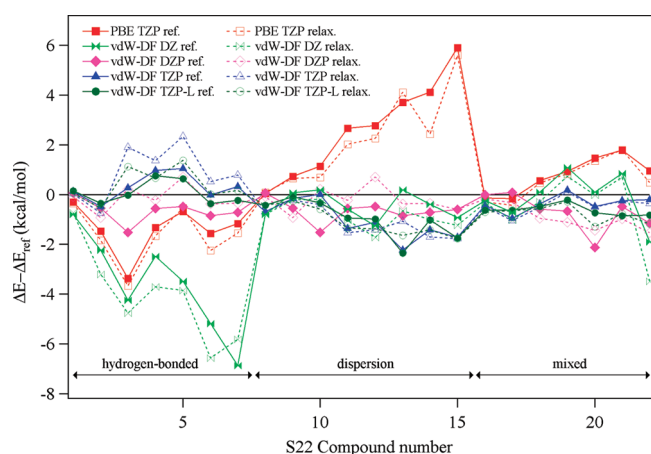


Figure 1. Binding energies differences ($\Delta E - \Delta E_{ref}$) for the S22 data set with respect to the S22A reference binding energies of Takatani *et al.*²⁹ The geometries of molecules are either fixed at the reference (ref.) S22 geometries of Jureka *et al.*²⁵ or fully relaxed (relax.).

Burns *et al.*,⁴⁰ in particular, only choosing results using a basis set (aug-cc-pVTZ) similar in size to the largest basis sets used in this study and choosing results without BSSE corrections.

The results in Table 2 highlight why using several error statistics can give a better gauge of the performance of a particular choice of XC functional and basis set, rather than just one. For example, if we only examined the MD of results, our vdW-DF TZP result for the full relaxed geometry gives the smallest deviation (-0.26 kcal/mol); however, when examining the MAD and RMSD results, we find it actually has slightly worse results than the DZP and TZP-L results for both the reference and fully relaxed geometries. Comparing the reference geometry results to the fully relaxed results we find that the errors are smaller for the results at the ideal reference geometries (by approximately 0.2 kcal/mol) when using the DZ, TZP, and TZP-L basis sets. Using the DZP basis sets, the results are very similar at the reference and fully relaxed geometries.

The overall results for the MAD and RMSD show that, when the vdW-DF functional is used, the DZP basis sets gives the best results, slightly outperforming the TZP-L basis set, which is slightly better than the TZP basis set results. Using the smallest DZ basis sets, the vdW-DF performs overall as badly, or worse, than the PBE functional, which was not designed to include van der Waals forces. Delving into the MAD and RMSD values in more detail, we can examine which basis set performs best for particular types of van der Waals complexes. For example, for the mixed complexes at the reference and fully relaxed geometries,

Table 2. Mean Deviation (MD), Mean Absolute Deviation (MAD), and Root Mean Square Deviation (RMSD) for S22 Data Set Binding Energies Using the S22A Reference Binding Energy Results of Takatani et al.^{29a}

	MD	MAD	RMSD
this study			
PBE TZP – ref.	0.76	1.68	2.23
vdW-DF DZ – ref.	–1.34	1.58	2.39
vdW-DF DZP – ref.	–0.53	0.58	0.78
vdW-DF TZP – ref.	–0.41	0.67	0.89
vdW-DF TZP-L – ref.	–0.53	0.67	0.85
PBE TZP – relax.	0.47	1.61	2.11
vdW-DF DZ relax.	–1.78	1.91	2.70
vdW-DF DZP relax.	–0.49	0.59	0.75
vdW-DF TZP relax.	–0.26	0.90	1.10
vdW-DF TZP-L relax.	–0.52	0.83	0.98
GGA-D			
B97-D3 ^b	0.14	0.36	0.44
BP86-D3 ^b	–0.73	0.77	0.96
hybrid functionals			
M05-2X ^b	0.42	0.64	0.85
M05-2X-D3 ^b	–0.28	0.39	0.49
PBE0-D3 ^b	–0.40	0.55	0.70
B3LYPD-D3 cc-pVTZ ^b	–0.32	0.37	0.49
wB97X-D cc-pVTZ ^b	–0.79	0.79	0.92
other			
SCS-MP2 ^c	–0.72	0.80	0.96
vdW-DF(revPBE) ^d	0.83	0.94	1.38
vdW-DF2(PW86) ^d	0.48	0.52	0.71

^a Binding energy error statistics are in kcal/mol and do not include BSSE corrections. Our results use either the reference (ref.) or fully relaxed (relax.) geometries. ^b Results from Burns et al.⁴⁰ using an aug-cc-pVTZ basis set. ^c Results from Takatani et al.²⁹ using an aug-cc-pVTZ basis set. ^d Results from Lee et al.²⁰

the TZP basis performs best. For hydrogen-bonded complexes, the TZP-L basis performs best at the reference geometry, and the DZP basis performs best at the fully relaxed geometry. Burns et al.⁴⁰ found that a DZP type basis performed slightly better for hydrogen bonded complexes, while a TZP type basis performed better for the dispersion complexes. These results illustrate how the choice of basis set and the type of intermolecular interaction affects the binding energy results.

Comparing the S22A error statistics in Table 2, we find that the vdW-DF in SIESTA performs well when compared to the subset chosen here of the many XC functional results reported by Burns et al.,⁴⁰ SCS-MP2 results of Takatani et al.,²⁹ and the vdW-DF/2 results of Lee et al.²⁰ In particular the MD, MAD, and RMSD for our vdW-DF calculations of the reference structures are similar in magnitude to the M05-2X and wB97X-D cc-pVTZ hybrid XC functionals and are slightly better than the SCS-MP2 and BP86-D3 results. The vdW-DF error statistics are slightly worse than the B97-D3 GGA functional results and the B3LYPD-D3 and M05-2X-D3s hybrid functional results. Lee et al.²⁰ report MAD and RMSD values of 0.94 and 1.38 kcal/mol, respectively, for their vdW-DF (revPBE) implementation. Our MAD and RMSD results from SIESTA are slightly better using the DZP, TZP, and

Table 3. Mean Deviation (MD), Mean Absolute Deviation (MAD), and Root Mean Square Deviation (RMSD) for S22 Data Set Binding Energies Using the S22 Reference Binding Energy Results of Jurecka et al.^{25a}

	MD	MAD	RMSD
this study			
PBE TZP – ref.	0.80	1.82	2.44
vdW-DF DZ – ref.	–1.30	1.60	2.48
vdW-DF DZP – ref.	–0.49	0.55	0.75
vdW-DF TZP – ref.	–0.36	0.56	0.73
vdW-DF TZP-L – ref.	–0.49	0.61	0.72
PBE TZP – relax.	0.51	1.74	2.32
vdW-DF DZ relax.	–1.74	1.87	2.77
vdW-DF DZP relax.	–0.44	0.53	0.64
vdW-DF TZP relax.	–0.21	0.75	0.94
vdW-DF TZP-L relax.	–0.48	0.73	0.83
GGA-D			
BLYP-D ^b	–0.39	0.53	0.64
B3LYP-D ^c	–0.28	0.48	
B97-D ^c	0.44	0.50	
hybrid functionals			
mPW2PLYP-D ^d	0.64	0.71	0.87
B2PLYP-D ^d	0.57	0.58	0.72
M05-2X ^e	0.29	0.57	0.72
M06-2X ^f	–0.14	0.44	0.56
XDM			
PW86PBE-XDM ^g		0.46	
other			
vdW-DF(revPBE) ^h	0.88	0.95	1.32
vdW-DF2(PW86) ^h	0.52	0.55	0.72
vdW-DF(revPBE) ⁱ	1.36	1.39	1.96
vdW-DF(PBE) ⁱ	–1.15	1.19	1.39
vdW-DF(revPBE) ^j		1.50	
vdW-DF(B86) ^j		0.53	
optB88-vdW ^j		0.23	

^a Binding energy error statistics are in kcal/mol and do not include BSSE corrections. Our results use either the S22 reference (ref.) or fully relaxed (relax.) geometries. ^b Results from Antony and Grimme³² using a TZV(2df,2pd) basis set. ^c Results from Chai and Head-Gordon¹¹ using a 6-311++G(3df,3pd) basis set. ^d Results from Schwabe and Grimme¹⁴ using a TZV(2df,2pd) basis set. ^e Results from Zhao and Truhlar³⁰ using a 6-31+G(d,p) basis set. ^f Results from Zhao and Truhlar⁴¹ using a 6-311+G(2df,2p) basis set. ^g Results from Kannemann and Becke³⁹ using the basis-set-free Numol method.⁴² ^h Results from Lee et al.²⁰ ⁱ Results from Gulans et al.⁴³ using a TZP basis set. ^j Results from Klimes et al.⁴⁴ using an aug-cc-pVTZ basis set.

TZP-L basis sets, both at the reference and the fully relaxed geometries, indicating that the performance of the vdW-DF in SIESTA is very good. Although the form of the vdW-DF (revPBE) functional is the same, these differences may arise due to the implementation in the respective codes, the type of basis set (e.g., planewave or localized orbitals), the quality of basis set, or the type of pseudopotentials used or other factors. Compared to the vdW-DF2 results of Lee et al.,²⁰ our results are slightly worse.

The newly revised S22A binding energies for the S22 data set were recently published by Sherrill and co-workers.^{29,40} S22

Table 4. Comparison of the Structural Parameters for *p*-*tert*-Butylcalix[4]arene · CS₂ from Theory and Experiment^a

	gas phase			solid state			
	vdW-DF	PBE ^b	HF ^c	ref.	relax.	ref.	exptl ^d
				vdW-DF	vdW-DF	PBE ^b	
<i>a</i> (Å)					12.911		12.770
<i>b</i> (Å)					12.811		12.770
<i>c</i> (Å)					13.185		13.314
Ar–O ₄ θ ₁ (deg)	120	126	125	120, 120	121, 121	125, 125	122
Ar–O ₄ θ ₂ (deg)	123	126	125	123, 123	125, 124	125, 125	122
Ar–O ₄ θ ₃ (deg)	121	126	125	124, 123	123, 124	124, 125	122
Ar–O ₄ θ ₄ (deg)	123	126	125	124, 124	124, 124	124, 124	122
C(CS ₂)–O ₄ (Å)	5.75	5.93	6.51	5.64, 5.66	5.58, 5.64	5.75, 5.80	5.45, 5.45
C–S (Å)	1.59, 1.61	1.59, 1.60	1.55	1.59, 1.61	1.59, 1.61	1.56, 1.58	1.50, 1.61
O _H –O _H (Å)	2.70, 2.71		2.73	2.69, 2.71	2.70, 2.71		2.68
O ₄ –CS ₂ (deg)	4.6	1.2	0.0	2.4, 2.5	2.9, 3.2	0.9, 1.2	0.0
I ^b (%)	89		57	89, 89	89, 90		92
<i>d</i> _A (Å)	8.20, 8.43		8.48	8.39, 8.49	8.40, 8.58		8.34

^aTheoretical structures are either fixed at the reference experimental lattice parameters (ref.) or have fully optimized lattice parameters and atomic coordinates (relax.). Ar–O₄ angles are the angles between the four aromatic rings (Ar) and the O₄ plane, which is defined by the four phenol oxygen atoms (O_H) in the calixarene. C(CS₂)–O₄ is the height of the CS₂ molecule above the O₄ plane. C–S is the CS₂ molecule bond lengths, with S₁ being the inner sulphur atom and S₂ being the outer atom. O₄–CS₂ is the tilt between the long axis of the CS₂ molecule and the O₄ plane. The distance *d*_A is the separation between the top-most carbon atoms of the aromatic rings (Ar) on opposite sides of the calixarene cage. The inclusion percentage (I^b) is calculated using $I^b = (\text{dist}(S_1-S_2) - \text{dist}(S_2-\text{plane}_b)) / (\text{dist}(S_1-S_2))$, where plane_b represents the top of the calixarene cavity. ^bResults from Ogden et al.⁴⁶ ^cResults from Schatz et al.⁴⁵ ^dCrystal structure from Schatz et al.³³

literature studies before these S22A results were published used the reference binding energies from Jurecka et al.²⁵ To enable direct comparison of our results to these earlier published studies, in Table 3 we compare the MD, MAD, and RMSD for our calculated binding energies based on the S22 binding energies reported by Jurecka et al.²⁵

The literature S22 binding error statistics in Table 3 are by no means an attempt to collate all studies in the computational literature; we have simply made a selection of the wide ranging literature with examples of different types of dispersion implementations such as XDM or hybrid methods for which to compare our results. Many of the studies^{14,32,35,40} we referenced in Table 3 reported results for other density functionals and also reported results for a variety of basis set sizes or other computational parameters.

Examining the error statistics in Table 3, in particular, the MAD and RMSD values, we find that at the reference geometries, our vdW-DF results compares well to other DFT-D methods. The MAD values of approximately 0.6 kcal/mol for the DZP, TZP, and TZP-L basis sets are similar to, or better than, the majority of literature results reported in Table 3. As mentioned before, using the fully relaxed geometries, the results are slightly worse for the DZ, TZP, and TZP-L basis sets. Comparing our results to the vdW-DF results of Lee et al.,²⁰ this time for the S22 binding energies, we see similar behavior to that for the S22A binding energies in Table 2 with our MAD and RMSD values being slightly better, both at the reference and the fully relaxed geometries. Again, our results are slightly worse compared to the vdW-DF2 results. Gulans et al.⁴³ reported their own implementations of vdW-DF(revPBE) and vdW-DF(PBE) in the SIESTA code, and their results appear to be much less accurate than our results at both the reference and fully relaxed geometries, although these values used a CP correction, which will alter these values,

particularly depending on their basis set choices etc. Klimes et al.⁴⁴ examined the vdW-DF based on six different GGA functionals, finding the MAD values varied from 1.50 kcal/mol for the vdWDF(revPBE) functional to 0.53 kcal/mol for the vdWDF-(B86) functional. At the reference geometries, our MAD values of 0.56 and 0.61 kcal/mol for the two basis set options indicate that the vdW-DF(revPBE) implementation in SIESTA performs extremely well compared to these results. Klimes et al.⁴⁴ also report MAD values for several new optimized functionals, with the optB88-vdW functional performing best with a MAD of 0.23 kcal/mol.

A particular advantage of SIESTA is that the localized numerical basis set enables us to examine (with full relaxation of unit cells and atomic coordinates) systems such as large biomolecules and soft matter systems, which would not be accessible to a planewave basis set or using high-order correlated based methods. The compromise for this speed and ability to examine large systems is accuracy; however, our results for the S22A and S22 binding energies in Tables 1–3 show that the vdW-DF implementation in SIESTA overall performs extremely well when using good quality DZP or TZP basis sets. In particular, when compared to vdW-DF implementations in other software codes reported by Lee et al.²⁰ and Klimes et al.,⁴⁴ the SIESTA MAD and RMSD values were better than or at least equally as good as the other implementations. A revised version of the vdW-DF has recently been published and is termed vdW-DF2²⁰ and reports improved accuracy for the S22 data set as reported above. In the future, if this is implemented in SIESTA, this could lead to even better performance for the S22 data set in SIESTA. There is little difference between the binding energies for the DZP, TZP, and TZP-L basis sets at the reference S22 geometries; however, when using fully relaxed geometries, the DZP basis set performs slightly better than the TZP-L and TZP basis sets. On the basis

of these results, we then examined the structure of two calixarene inclusion compounds.

B. *p*-tert-Butylcalix[4]arene·CS₂ and *p*-tert-Butylcalix[4]arene·toluene. We have performed vdW-DF calculations on the *p*-tert-butylcalix[4]arene structure with toluene and carbon disulfide (CS₂) guest molecules to examine the performance of the vdW-DF for calixarene host–guest structures. In Table 4, we report the structural parameters of the *p*-tert-butylcalix[4]arene·CS₂ compound in both the gas phase and the solid state. We examined the solid state calixarene structure fixed at the experimental lattice parameters of Schatz et al.³³ (allowing atomic coordinates to relax), and with fully optimized lattice parameters and atomic coordinates, to allow for comparison with other theoretical investigations.^{45,46}

The fully optimized *p*-tert-butylcalix[4]arene·CS₂ structure in Table 4 shows good agreement with the room temperature experimental crystal structure of Schatz et al.³³ For the fully optimized crystal structure, there is a small distortion from a tetragonal to orthorhombic unit cell, although the lattice parameters are within 1% of the experimental values, and the rest of the structural properties are in good agreement with those calculated at the experimental lattice parameters.

Comparing the values in Table 4, the main differences that appear in both the gas phase and solid state are the C(CS₂)–O₄ distance and the inclusion percentage (*I*^b), both of which give a measure of how much the CS₂ guest molecule is included within the calixarene cage (a higher inclusion percentage will have a smaller C(CS₂)–O₄ distance and vice versa). For both *I*^b and C(CS₂)–O₄ values, the vdW-DF results in the gas phase and solid state phase are much closer to the experimental values than the other theoretical calculations.

The inclusion percentage for the vdW-DF calculations is approximately 90% in both the gas phase and solid state, very close to the experimental value of 92% and noticeably higher than the HF value of 57%. Correspondingly, the C(CS₂)–O₄ distances are shorter in the vdW-DF calculations with 5.75 Å in the gas phase to 5.58–5.66 Å in the solid state (reference and relaxed geometries, respectively) for the vdW-DF. The PBE results for the C(CS₂)–O₄ distances are 5.93 and 5.75–5.80 Å for the gas phase and solid state, respectively, and 6.51 Å for the gas phase HF results calculations. In theoretical calculations using the vdW-DF and PBE functionals, there is a small tilt of the CS₂ molecule of several degrees with respect to the O₄ plane, which is not seen in the experimental structure or HF calculations. In Figure 2a, we show the gas phase (light gray) and fully relaxed solid state structures (dark gray) of *p*-tert-butylcalix[4]arene·CS₂, showing a single calixarene unit overlaid for each. In this figure, the CS₂ molecule is clearly further into the cage in the solid state (dark gray) structure than the gas phase structure (light gray), as quantified earlier by the C(CS₂)–O₄ distances. There are also some small differences in the orientations of the methyl side chains of the *tert*-butyl groups.

In Table 5, we report the structural parameters of the *p*-tert-butylcalix[4]arene·toluene compound in both the gas phase and the solid state. We examined the solid state calixarene structure fixed at the experimental *P*112/*a* crystal structure reported by Arduini et al.³⁴ (allowing atomic coordinates to relax) and with fully optimized lattice parameters and atomic coordinates.

The experimental crystal structure of *p*-tert-butylcalix[4]arene·toluene has been reported in several different crystal structures at a range of temperatures. The high temperature structures were reported as being in the tetragonal *P*4/*n* space

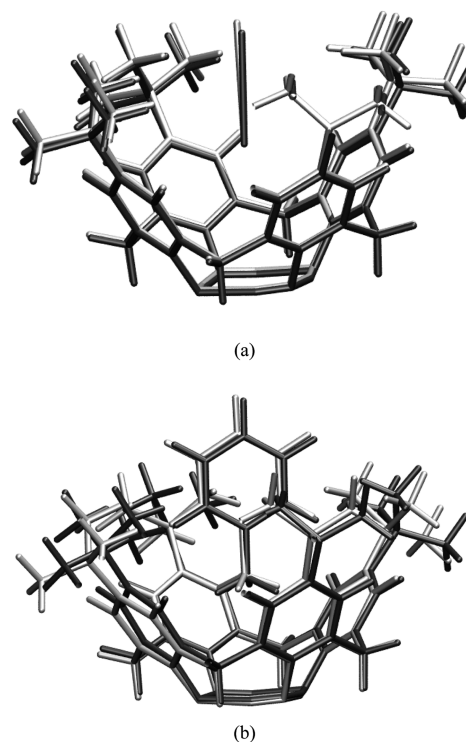


Figure 2. Superposition of the gas phase and solid state (fully relaxed) structures of (a) *p*-tert-butylcalix[4]arene·CS₂ and (b) *p*-tert-butylcalix[4]arene·toluene. The gas phase structures and solid state crystal structures are colored light gray and dark gray, respectively.

group;⁴⁷ then later structural refinements reported monoclinic structures.^{34,48} In Table 5, we report the two monoclinic structures for comparison to our computational results. In the monoclinic structures, the toluene guest molecules induce a distortion in the host calixarene molecules, and there is a correlation of guest molecules in adjacent calixarene molecules.⁴⁸ If no distortion occurred, the four Ar–O₄ angles in Table 5 would be the same value. The actual position of the toluene molecule in the experimental structures is dynamically disordered and typically averages to give either a 4-fold⁴⁷ or 2-fold^{34,48} symmetry. In this study, we perform static electronic structure energy minimizations. Molecular dynamics simulations would be required to investigate the dynamical nature of the guest molecule, but this was not the focus of this investigation.

Examining the gas phase structures in Table 5, we find that the vdW-DF structure is slightly closer to the experimental structures than the PBE results of Ogden et al.,⁴⁶ in particular for the Ar–O₄ angles and the CH₃–O₄ distance. The CH₃–O₄ distance is a measure of how much the toluene molecule is included within the host molecule, so the PBE value of 3.93 Å, compared to the vdW-DF value of 3.77 Å, indicates that the toluene is more included in the vdW-DF calculations.

Looking now at the solid state structures there is quite reasonable agreement between the vdW-DF and PBE results⁴⁶ with the experimental structures. For the vdW-DF results at the experimental (ref.) geometry, the CH₃–O₄ distances are 3.61 and 3.66 Å, very similar to the PBE results of 3.64 and 3.66 Å. The experimental monoclinic structures report a similar lower value in the range of 3.65–3.70 Å, but the second distance is about 0.1 Å longer than the vdW-DF and PBE results. The interplanar tilt of the toluene molecules is also slightly larger in the vdW-DF

Table 5. Comparison of the Structural Parameters for *p*-*tert*-Butylcalix[4]arene · toluene from Theory and Experiment^a

	gas phase		solid state				
	vdW-DF	PBE ^b	ref.	relaxed	ref.		
			vdW-DF	vdW-DF	PBE ^b	exptl. ^c	exptl. ^d
<i>a</i> (Å)				18.239	12.756	17.889	17.814
<i>b</i> (Å)				18.127	12.756	17.899	17.806
<i>c</i> (Å)				13.647	13.793	13.827	13.890
Ar–O ₄ θ_1 (deg)	118	123	117, 119	119, 120	123, 121	119	120
Ar–O ₄ θ_2 (deg)	124	132	124, 122	125, 125	127, 128	124	125
Ar–O ₄ θ_3 (deg)	118	124	120, 120	121, 121	122, 123	120	120
Ar–O ₄ θ_4 (deg)	125	127	126, 126	125, 124	126, 126	125	125
CH ₃ –O ₄ (Å)	3.77	3.93	3.61, 3.66	3.63, 3.67	3.64, 3.66	3.70, 3.84	3.64, 3.83
toluene tilt (deg)	0.3	3.5	3.7, 4.1	4.2, 5.5	1.1, 3.8	0	4.2, 7.7
interplanar angle	19.1		23.7, 26.6	23.3, 26.3	19.0, 24.8	19.6, 23.1	18.0, 21.9

^aTheoretical structures are either fixed at the reference experimental lattice parameters (ref.) or have fully optimized lattice parameters and atomic coordinates (relax.). Ar–O₄ angles are the angles between the four aromatic rings (Ar) and the O₄ plane, which is defined by the four phenol oxygen atoms (O_H) in the calixarene. CH₃–O₄ is the height of the toluene molecule above the O₄ plane. The toluene tilt is the angle between the long axis of the toluene molecule and the O₄ plane. The interplanar angle is the angle between the pseudo-mirror plane of two of the methylene C atoms and the plane of the aromatic ring in the toluene molecule. ^bResults from Ogden et al.⁴⁶ using the *P4/n* crystal structure of Andreotti et al.,⁴⁷ where $\alpha = \beta = \gamma = 90^\circ$. ^cCrystal structure from Arduini et al.³⁴ using the *P112/a* space group. ^dCrystal structure from Enright et al.⁴⁸ using the *P2/c* space group, where $\alpha = \beta = 90$ and $\gamma = 89.91^\circ$.

Table 6. Binding Energy (kcal/mol) per Guest Molecule of *p*-*tert*-Butylcalix[4]arene · toluene (Toluene) and *p*-*tert*-Butylcalix[4]arene · CS₂ (CS₂) for the Gas Phase and Solid State Structures^a

	gas phase		
	vdW-DF	PBE DZP ^b	
CS ₂	–13.27	–0.76	
toluene	–21.78		
	solid state		
	vdW-DF		PBE DZP ^b
	fixed at exptl unit cell	relaxed unitcell	fixed at exptl unit cell
CS ₂	–83.42	–83.59	–1.84
toluene	–92.27	–92.67	

^aThe solid state structures are those reported in Tables 4 and 5. ^bResults from Ogden et al.⁴⁶ using the *P4/n* crystal structure of Andreotti et al.,⁴⁷ where $\alpha = \beta = \gamma = 90^\circ$.

calculation, with angles of 23.7 and 26.6°, compared to experimental values ranging from 18 to 23°. When we fully relaxed the solid state structure, we found that the lattice parameter changes were less than 2% different compared to the experimental crystal structure. The CH₃–O₄ distances of 3.63 and 3.67 Å are similar to the result fixed at the experimental lattice parameters and are again slightly less than the experimental results. The interplanar angles of 23.3 and 26.3° are slightly larger than for the fixed lattice parameters (values of 23.7 and 26.1°) but are slightly overestimated compared with the experimental values. Overall, the PBE and vdW-DF functionals appear to perform quite similarly for the structural parameters of the solid state structures, while the vdW-DF appears to perform slightly better for the gas phase

structure. In Figure 2b, we show the gas phase (light gray) and fully relaxed solid state structures (dark gray) of *p*-*tert*-butylcalix[4]arene · toluene, showing a single calixarene unit overlaid for each. In this figure, the toluene molecule sits further into the cage in the solid state (dark gray) structure than the gas phase structure (light gray), as quantified earlier by the CH₃–O₄ distances. There are also some small differences in the orientations of the methyl side chains of the *tert*-butyl groups.

After examining the structural properties of the *p*-*tert*-butylcalix[4]arene · CS₂ and *p*-*tert*-butylcalix[4]arene · toluene compounds, we now examine the binding energies. In Table 6, we compare the binding energies of the *p*-*tert*-butylcalix[4]arene · toluene and *p*-*tert*-butylcalix[4]arene · CS₂ compounds in both the gas phase and solid state, defining the binding energy as per guest molecule, with reference to the isolated gas phase species. As reported earlier with the S22 data set, we do not include any BSSE corrections in these binding energies.

The binding energy results in Table 6 clearly show that the binding of toluene and CS₂ molecules is weaker in the gas phase than the solid state. This is not unsurprising as the cavity in the solid state is well-defined and the guest molecules are also influenced by binding with neighboring calixarene host molecules. For the *p*-*tert*-butylcalix[4]arene · CS₂ structure, the binding energy in the gas phase is –13.27 kcal/mol, while in the solid state, the values are –83.42 and –83.59 kcal/mol, respectively, for the fixed and fully relaxed structures, respectively. For the *p*-*tert*-butylcalix[4]arene · toluene in the gas phase, the binding energy is –21.78 kcal/mol, compared to –92.27 and –92.67 kcal/mol for the fixed and fully relaxed solid state structures, respectively. Previous PBE calculations of Ogden et al.⁴⁶ of the *p*-*tert*-butylcalix[4]arene · CS₂ structure also predicted this behavior with values of –0.76 and –1.84 kcal/mol for the gas phase and solid state structures, respectively, although they acknowledge the magnitude of the binding energies will be underestimated due to a lack of suitable description of the dispersion forces. Overall, the vdW-DF in SIESTA appears to provide good

descriptions of both the gas phase and solid state structures of *p*-*tert*-butylcalix[4]arene·CS₂ and *p*-*tert*-butylcalix[4]arene·toluene, based on structural properties reported in Tables 4 and 5 and the binding energies reported in Table 6 and in most cases provide superior results to other theoretical studies.

CONCLUSIONS

We have investigated the performance of the vdW-DF implemented in the SIESTA code for the S22 data set, examining the effect of basis set choice and atomic relaxation. Using the MD, MAD, and RMSD error statistics, we have quantified the results against both the S22²⁵ and S22A³⁹ reference binding energies and find that SIESTA performs very well compared to a range of results from other studies. We find that at the reference geometries there is little difference between the DZP, TZP, and TZP-L basis sets; however, when full atomic relaxations are carried out, the DZP basis overall gives the best results by approximately 0.2 kcal/mol. Dividing the S22 compounds into their dominant van der Waals interactions (hydrogen bonded, dispersion dominated, or mixed), we also demonstrated how different basis sets performed better for particular types of van der Waals interactions.

We then examined the performance of vdW-DF for two calixarene host–guest inclusion compounds, namely, *p*-*tert*-butylcalix[4]arene·CS₂ and *p*-*tert*-butylcalix[4]arene·toluene. We examined both the gas phase and solid state structures and compared the results against other theoretical investigations and experimental data. Overall, for both inclusion compounds, vdW-DF performs extremely well and outperforms other Hartree–Fock and DFT PBE results where published. In particular, for the *p*-*tert*-butylcalix[4]arene·CS₂ compound, the vdW-DF calculated structures show that the CS₂ molecule is included further into the host compound than other theoretical calculations and much better matches the experimental structures. Binding energy calculations show that the guest molecules are much more strongly bound in the solid state than the gas phase, as expected.

AUTHOR INFORMATION

Corresponding Author

*E-mail: a.rohl@curtin.edu.au

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

The authors thank iVEC and the NCI for the provision of computational resources. A complete archive of the SIESTA calculations of the S22 data set is available at <http://dx.doi.org/10.4225/06/4ED6B979EBEC4>.

REFERENCES

- (1) Soler, J. M.; Artacho, E.; Gale, J. D.; Garcia, A.; Junquera, J.; Ordejon, P.; Sanchez-Portal, D. *J. Phys.: Condens. Matter* **2002**, *14*, 2745–2779.
- (2) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479–483.
- (3) Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1463–1473.
- (4) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787–1799.
- (5) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *J. Chem. Phys.* **2010**, *132*, 154104.
- (6) Williams, R. W.; Malhotra, D. *Chem. Phys.* **2006**, *327*, 54–62.

- (7) Riley, K. E.; Vondrasek, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2007**, *9*, 5555–5560.
- (8) Civalleri, B.; Zicovich-Wilson, C. M.; Valenzano, L.; Ugliengo, P. *CrystEngComm* **2008**, *10*, 405–410.
- (9) Neumann, M. A.; Leusen, F. J. J.; Kendrick, J. *Angew. Chem., Int. Ed.* **2008**, *47*, 2427–2430.
- (10) Xu, X.; Goddard, W. A., III. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 2673–2677.
- (11) Chai, J.-D.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615–6620.
- (12) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2006**, *120*, 215–241.
- (13) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 5656–5667.
- (14) Schwabe, T.; Grimme, S. *Phys. Chem. Chem. Phys.* **2007**, *9*, 3397–3406.
- (15) Zhang, Y.; Xu, X.; Goddard, W. A., III. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 4963–4968.
- (16) Hebelmann, A. *J. Chem. Phys.* **2009**, *130*, 084104.
- (17) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2005**, *122*, 154104.
- (18) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2006**, *124*, 014104.
- (19) Dion, M.; Rydberg, H.; Schroder, E.; Langreth, D. C.; Lundqvist, B. I. *Phys. Rev. Lett.* **2004**, *92*, 246401.
- (20) Lee, K.; Murray, E. D.; Kong, L.; Lundqvist, B. I.; Langreth, D. C. *Phys. Rev. B* **2010**, *82*, 081101.
- (21) Vydrov, O. A.; Van Voorhis, T. *Phys. Rev. Lett.* **2009**, *103*, 063004.
- (22) Roman-Perez, G.; Soler, J. M. *Phys. Rev. Lett.* **2009**, *103*, 096102.
- (23) Kong, L.; Roman-Perez, G.; Soler, J. M.; Langreth, D. C. *Phys. Rev. Lett.* **2009**, *103*, 096103.
- (24) Walker, A. M.; Civalleri, B.; Slater, B.; Mellot-Draznieks, C.; Cora, F.; Zicovich-Wilson, C. M.; Roman-Perez, G.; Soler, J. M.; Gale, J. D. *Angew. Chem., Int. Ed.* **2010**, *122*, 7663–7665.
- (25) Jurecka, P.; Sponer, J.; Cerny, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985–1993.
- (26) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- (27) Troullier, N.; Martins, J. L. *Phys. Rev. B* **1991**, *43*, 1993–2006.
- (28) Junquera, J.; Paz, O.; Sanchez-Portal, D.; Artacho, E. *Phys. Rev. B* **2001**, *64*, 235111.
- (29) Takatani, T.; Hohenstein, E. G.; Malagoli, M.; Marshall, M. S.; Sherrill, C. D. *J. Chem. Phys.* **2010**, *132*, 144104.
- (30) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 289–300.
- (31) van Mourik, T.; Karamertzanis, P. G.; Price, S. L. *J. Phys. Chem. A* **2006**, *110*, 8–12.
- (32) Antony, J.; Grimme, S. *Phys. Chem. Chem. Phys.* **2006**, *8*, 5287–5293.
- (33) Schatz, J.; Scholdbach, F.; Lentz, A.; Rastatter, S.; Schilling, J.; Dormann, J.; Ruoff, A.; Debaerdemaecker, T. *Z. Naturforsch., B: J. Chem. Sci.* **2000**, *55*, 213–221.
- (34) Arduini, A.; Caciuffo, R.; Geremia, S.; Ferrero, C.; Ugozzoli, F.; Zontone, F. *Supramol. Chem.* **1998**, *10*, 125–132.
- (35) Jurecka, P.; Cerny, J.; Hobza, P.; Salahub, D. R. *J. Comput. Chem.* **2007**, *28*, 555–569.
- (36) Koch, W.; Holthausen, M. C. *A Chemist's Guide to Density Functional Theory*, 2nd ed.; Wiley-VCH Verlag GmbH: Weinheim, Germany, 2001.
- (37) Carter, D. J.; Rohl, A. L.; Gale, J. D. *J. Chem. Theory Comput.* **2006**, *2*, 797–800.
- (38) Carter, D. J.; Rohl, A. L. *J. Chem. Theory Comput.* **2011**, *7*, 1604–1609.
- (39) Kannemann, F. O.; Becke, A. D. *J. Chem. Theory Comput.* **2010**, *6*, 1081–1088.
- (40) Burns, L. A.; Vazquez-Mayagoitia, A.; Sumpter, B. G.; Sherrill, C. D. *J. Chem. Phys.* **2011**, *134*, 084107.
- (41) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. C* **2008**, *112*, 4061–4067.
- (42) Becke, A. D.; Dickson, R. M. *J. Chem. Phys.* **1998**, *89*, 2993–2997.
- (43) Gulans, A.; Puska, M. J.; Nieminen, R. M. *Phys. Rev. B* **2009**, *79*, 201105.

- (44) Klimes, J.; Bowler, D. R.; Michaelides, A. *J. Phys.: Condens. Matter* **2010**, *22*, 022201.
- (45) Schatz, J.; Backes, A. C.; Siehl, H. *J. Chem. Soc., Perkin Trans. 2* **2000**, *2*, 609–610.
- (46) Ogden, M. I.; Rohl, A. L.; Gale, J. D. *Chem. Commun.* **2001**, 1626–1627.
- (47) Andreeti, G. D.; Ungaro, R.; Pochini, A. *J. Chem. Soc., Chem. Commun.* **1979**, 1005–1007.
- (48) Enright, G. D.; Brouwer, E. B.; Udachin, K. A.; Ratcliffe, C. I.; Ripmeester, J. A. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *58*, 1032–1035.

The Fast-Folding Mechanism of Villin Headpiece Subdomain Studied by Multiscale Distributed Computing

Ryuhei Harada^{†,‡,§} and Akio Kitao^{*,†,‡,§}[†]Department of Physics, Graduate School of Science, The University of Tokyo, Tokyo, 7-3-1, Hongo, Bunkyo-ku 113-0033, Japan[‡]Institute of Molecular and Cellular Bioscience, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-0032, Japan[§]Japan Science and Technology Agency, Core Research for Evolutional Science and Technology, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-0032, Japan

ABSTRACT: The fast-folding mechanism of a 35-residue mini-protein, villin headpiece subdomain (HP35), was investigated using folding free energy landscape analysis with the multiscale free energy landscape calculation method (MSFEL). A major and a minor folding pathway were deduced from the folding free energy landscape. In the major folding pathway, the formation of helices II and III was the rate-limiting step in the transition to an intermediate state, triggered by the folding of the PLWK motif. HP35 then folds into the native structure through the formation of the hydrophobic core located at the center of the three-helix bundle. Mutations in the motif and hydrophobic core that suppressed folding into the native state drastically changed the folding free energy landscape compared to the wild type protein. In the minor folding pathway, nucleation of the hydrophobic core preceded formation of the motif.

1. INTRODUCTION

The folding of proteins into their native structures plays an important role in many biological processes. Protein misfolding and aggregation into amyloid-like fibrils are thought to cause some diseases.^{1–5} Recent dramatic advances in the power of computational methods have enabled researchers to trace a series of folding pathways from molecular dynamics (MD) simulations on the order of microsecond time scales. Significant progress has been made recently in *ab initio* folding simulations of some fast-folding mini-proteins, including chignolin (10 residues),^{6,7} Trp-cage (20 residues),^{8,9} and HP35 (35 residues).^{10–13} Predicted structures of small proteins typically deviate from experimental structures determined by NMR or X-ray crystallography by 2–4 Å in the C α atoms. Conventional MD (CMD) allows for direct observation of the time course of folding events. Information obtained in the form of MD trajectories provides dynamic pictures of conformational transitions. However, generating multiple trajectories sufficient for establishing kinetic views of protein folding is challenging.

The free energy landscape (FEL) introduces an important concept for investigating free energy changes on a given reaction coordinate space from a statistical point of view. However, the complex energy surface prevents accurate conformational sampling of the FEL due to trappings into local energy minima, resulting in bad conformational samplings. Therefore, samplings of rare events like conformational transitions jumping among minima are important to calculate the FEL accurately. To overcome this sampling problem, many computational methodologies have been developed, including the extended ensemble method like the multicanonical MD (McMD)^{14,15} and replica exchange molecular dynamics (REMD)^{16,17} methods. In the McMD method, a non-Boltzmann sampling enables random walks on the energy space without trapping in local energy minima. A target ensemble can be reconstructed by reweighting. In the REMD, a set of simulations is performed at different temperatures, and temperature exchanges are periodically attempted according to the Metropolis criterion,¹⁸ which attains

random walks in the temperature space and prompt escape from local energy minima. The combination of the REMD with the umbrella sampling like REUS (replica-exchange umbrella sampling)¹⁹ and bias-exchange method²⁰ is also an effective approach to enhancing conformational sampling. In these methods, the umbrella potentials are exchanged in the REMD as well as temperatures. Metadynamics²¹ is a powerful method that can be used both for calculating free energy and for accelerating rare events in systems. In this method, the normal evolution of the system is biased by a history-dependent potential constructed as a sum of Gaussians centered along the trajectory followed by a suitably chosen set of collective variables. The sum of Gaussians is used for reconstructing iteratively as an estimator of the free energy and forcing the system to escape from local minima. The Wang–Landau method²² is an extended Monte Carlo method for calculating the density of state efficiently by performing independent random walks in different and restricted ranges of energy. The resultant density of states is modified continuously to produce locally flat histograms. This method permits us to directly access the free energy and entropy. The transition path sampling method²³ is a method for sampling a conformational transition called the “reactive path” that connects a given reactant and product conformation starting from an arbitrary initial transition path. The reactive transitional paths are sampled with the Metropolis criteria so as to hold a certain ensemble. This method has been successful in folding studies and conformational samplings of small proteins.^{24–27} Transform and relax sampling (TRS) was successful in sampling protein domain motion and mini-protein folding.²⁸ Recently, we proposed a new approach to calculating the FEL, called the multiscale free energy calculation method (MSFEL).²⁹ In this method, multiple conformations are generated to cover a broad conformational space with a coarse-grained (CG) model. Distributed all-atom (AA)

Received: May 31, 2011

Published: December 08, 2011

MD simulations are then performed using umbrella sampling^{30,31} to sample local energy landscapes in parallel. Finally, the FEL is calculated using the weighted histogram analysis method (WHAM).^{32–34} The MSFEL method has been applied to the study of short peptides and mini-proteins, and the efficiency of the FEL calculation has been demonstrated.^{29,35}

In this study, we investigated the fast-folding mechanism of the 35 amino acid residue villin headpiece subdomain (HP35) in explicit solvent using the MSFEL method. HP35 is an F-actin-binding domain located on the far C-terminus of the super villin.^{36,37} HP35 can spontaneously fold into its native structure within microseconds without the assistance of disulfide bonds and metal ions. The native structure of HP35 has been determined using both NMR and X-ray crystallography in high resolution.^{38,39} Since HP35 is small in size and folds quickly and cooperatively, it has also been studied extensively using kinetic experiments,^{40,41} mutagenesis,^{42,43} and computer simulations.^{10–13} The folding FEL of HP35 by computer simulations has been first investigated in implicit solvent,¹² and intermediate and transitional conformations on the folding pathway have been reported.¹³ The folding of HP35 has been also investigated in explicit solvent.^{44–49} In experiments, mutational analyses of key residues have revealed the mechanism of fast folding.^{50,51}

For further understanding of the folding mechanism of HP35, especially, the information of the folding FEL in explicit solvent, which is still difficult to measure by experimentation, is an important factor in the filling of gaps between experiments and computations. The accurate estimations of free energy difference from the folding FEL calculations are quite meaningful for comparing and supporting experimental data in addition to the elucidation of the folding pathway. Therefore, we focused on the folding FEL of HP35 in explicit solvent and more accurately calculated the folding FEL using the MSFEL.

2. MATERIAL AND METHODS

Implementation of the MSFEL. A MSFEL analysis consists of four stages. In the first stage, a CG MD simulation is performed to sample a broad conformational space. To address the folding of HP35, a replica exchange MD (REMD) simulation¹⁷ with a CG model was used to further enhance the conformational sampling. The CG REMD simulations were performed by our original MD program developed for the MSFEL. We employed a C_α -based CG model²⁹ to widely sample the conformational space around a reference structure. The potential energy function was defined as the sum of bond, angle, torsion, and Lennard-Jones-type energy terms as follows:

$$\begin{aligned}
 V^{\text{CG}}(\vec{r}^{\text{C}\alpha} | \vec{r}^{\text{C}\alpha 0}) = & \sum_{|i-j|=1, i < j} k_{12} (r_{ij}^{\text{C}\alpha} - r_{ij}^{\text{C}\alpha 0})^2 \\
 & + \sum_{|i-j|=2, i < j} k_{13} (r_{ij}^{\text{C}\alpha} - r_{ij}^{\text{C}\alpha 0})^2 \\
 & + \sum_{|i-j|=3, i < j} k_{14} (r_{ij}^{\text{C}\alpha} - r_{ij}^{\text{C}\alpha 0})^2 \\
 & + \sum_{\substack{|i-j| < r_c, \\ |i-j| > 3}} k_{\text{LJ}} \left[\left(\frac{r_{ij}^{\text{C}\alpha 0}}{r_{ij}^{\text{C}\alpha}} \right)^{12} - \left(\frac{r_{ij}^{\text{C}\alpha 0}}{r_{ij}^{\text{C}\alpha}} \right)^6 \right]
 \end{aligned} \quad (1)$$

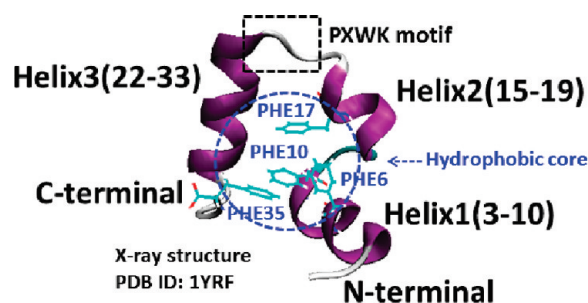


Figure 1. The native structure of HP35. The hydrophobic core residues, PLWK motif, and the definition of the two segments are shown. Figure created using VMD.⁶⁷

where $r_{ij}^{\text{C}\alpha 0}$ and $r_{ij}^{\text{C}\alpha}$ represent the distance between the i th and j th atoms of the reference and instantaneous structures, respectively. The fourth term is associated with the atom pairs within the cutoff distance $r_c = 10.0 \text{ \AA}$ in the native structure. To build the CG potential, the reference C_α coordinates were adopted from the X-ray structure of HP35 (Protein Data Bank ID: 1YRF)³⁸ shown in Figure 1 and were also used as the initial structures of the CG REMD. The global energy minimum of this potential function is designed to be the X-ray structure. Therefore, this CG model can be considered a Go-like model, a type of model widely used in protein folding studies.^{52–54} The parameter $r_{ij}^{\text{C}\alpha 0} = 3.8 \text{ \AA}$ was determined first to reproduce the optimal distance between two adjacent C_α atoms. The value of k_{12} was determined to best reproduce the distribution with the relation between the variance of the Gaussian at 300 K, $\sigma = 5.651 \times 10^{-2} \text{ \AA}$, $\beta = 5.919 \times 10^{-1} \text{ kcal/mol}$, and $k_{12} = 1/\beta\sigma^2$. The other parameters were defined by the ratios to k_{12} as $k_{13}/k_{12} = 1/5$ and $k_{14}/k_{12} = k_{\text{LJ}}/k_{12} = 1/100$, and the Newtonian equation of motion is integrated by a time step 15 ps. The CG REMD simulations were performed with 10 replicas at exponentially distributed temperatures of 200, 239, 286, 342, 404, 489, 585, 700, 853, and 1041 K. A 10^6 -step production CG MD run after a 10^3 -step equilibration was performed under a canonical ensemble with a Nosé–Hoover chain thermostat.⁵⁵ The two replicas with neighboring temperatures were exchanged every 10^3 steps, and a total of 10^4 snapshots were recorded every 10^2 steps. The average exchange rate between replicas was 0.31, which is sufficiently high to achieve efficient sampling.

In the second stage, multiple representative structures were chosen. These structures should roughly cover the entire conformational space sampled in the CG MD analysis and be distributed densely enough so that each AA MD trajectory in the third stage significantly overlaps with its neighboring trajectories. From structures obtained using the CG model, BBQ⁵⁶ was used to generate main-chain atoms from the C_α coordinates. Next, SCWRL⁵⁷ was employed to generate side-chain atoms. A total of 100 AA structures were constructed in this study. The C_α coordinates were picked up from 10 REMD trajectories with equal intervals.

In the third stage, independent AA MD simulations were conducted to investigate each local FEL more accurately around the distributed initial structures. The 100 AA structures were solvated in a rectangular box ($63.9 \text{ \AA} \times 55.2 \text{ \AA} \times 48.1 \text{ \AA}$) containing 3456 TIP3P water molecules.⁵⁸ Two chloride ions were also added to neutralize the system. The AA MD simulations were independently performed using the PMEMD module of the Amber 9.0 software⁵⁹ with the Amber parameter ff03 force field.⁶⁰ For the purpose of intensive local sampling, we employed the umbrella sampling method^{30,31} and used harmonic positional restraints for the C_α atoms as the umbrella potentials.²⁹ Short

energy minimizations and 300 ps relaxation MD runs that included density adjustments with an isothermal–isobaric ensemble at 300 K and 1 bar were then conducted using the Berendsen method. The systems were equilibrated with a canonical ensemble for 100 ps with harmonic restraints (1.0×10^{-4} kcal/mol/Å²) imposed on the C_{α} atoms (except for the N- and C-terminal residues) for umbrella samplings. Production runs were performed for 1 ns \times 100 trajectories, and each trajectory was recorded every 1.0 ps. In both the equilibration and production runs, the temperature was maintained at 300 K, and the SHAKE algorithm⁶¹ was used to enable the use of a long time step of 2 fs. Electrostatic interactions were calculated using the particle-mesh Ewald method⁶² with a real space cutoff distance of 9 Å.

In the final stage, the probability distributions obtained in the previous stage were reweighted and combined using the WHAM.^{32–34}

3. RESULTS AND DISCUSSION

3.1. Overview of the MSFEL. In the MSFEL method, the CG MD simulation is performed in the first stage to efficiently sample the conformational space. Figure 2 shows the time series of the C_{α} root-mean square deviation (C_{α} -RMSD) of two segments defined as described previously.^{12,13} The A and B segments are composed of helices I–II (residues 3–21) and helices II–III (residues 15–33), respectively. Segments A and B overlap with helix II in order to consider local folding between helices I and III with respect to helix II. As shown in Figure 2a, both segments folded and unfolded frequently during the CG MD simulation, which is indicative of efficient conformational sampling. For comparison, we performed a conventional long (100 ns) AA MD simulation at 300 K in explicit solvent after a 1 ns equilibration starting from the native structure (Figure 2b). In the CMD simulation, HP35 was trapped around the native state, resulting in insufficient conformational sampling for calculating an accurate folding FEL. These results indicate that the CG MD simulation enables enhanced conformational searching with relatively low computational cost. We also examined the growth of the three helices, helix I (residues 3–10), helix II (15–19), and helix III (22–33) in the CG MD. We defined an order parameter Φ as the ratio of the helical residues in each helix. For each snapshot of the CG MD, main-chain atoms were generated by BBQ,⁵⁶ and the secondary structure assignment was done by STRIDE.⁶³ As shown in the time of evolution of Φ (Figures 2c–e), folding and unfolding of the helices were frequently observed.

In the second stage, 100 CG structures were selected from trajectories of CG REMD simulations. From each trajectory that contains 10^4 snapshots of the C_{α} coordinates, 10 snapshots are selected at every 10^3 intervals (10 snapshots \times 10 replicas). Then, 100 AA structures were generated from the C_{α} coordinates using CG-AA mappings. The projections of the CG REMD trajectories onto the subspace spanned by C_{α} -RMSDs for segments A and B of the native structure (Figure 3a) and those of the 100 selected CG structures (Figure 3b) shows that the representative structures roughly cover the conformational space sampled in the CG level. The overlaps of the AA MD trajectories with its neighboring trajectories (Figure 3c) also indicate that the representative structures were dense enough (also see Figures 5a–c later). Since accuracy in CG-AA mapping is important, mappings were examined with a benchmark coordinate set similar to that used in our previous work.²⁹ A total of 10 000 AA coordinates that included both native and unfolded structures were generated using AA MD. Next, C_{α} coordinates were picked, and AA

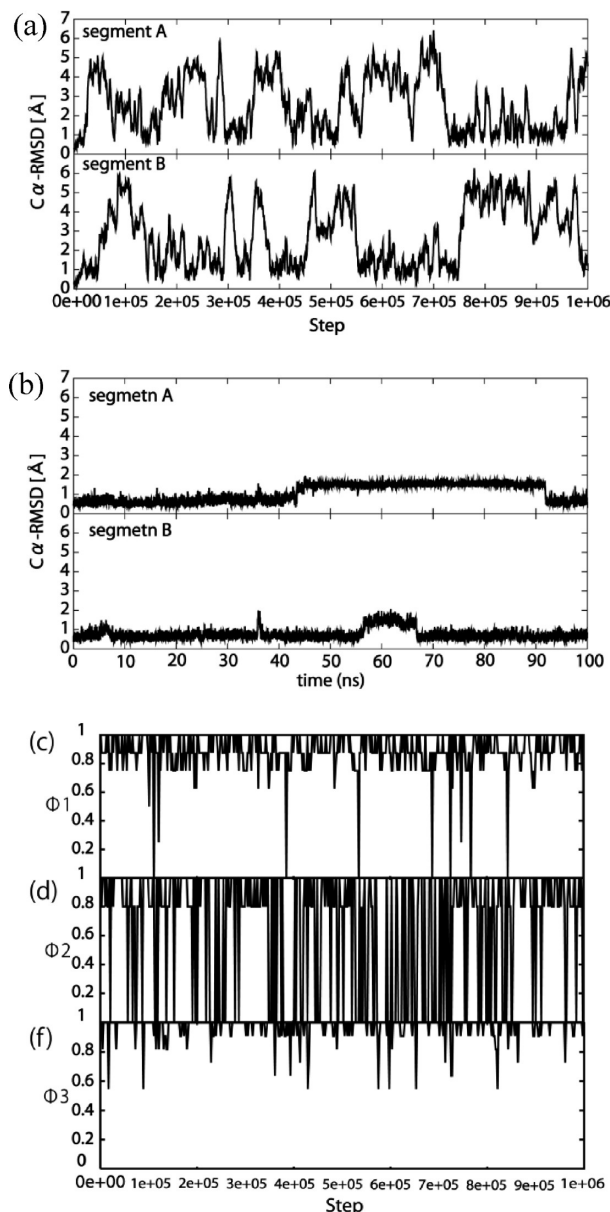


Figure 2. (a) Time series of the C_{α} -RMSD of segment A (helices I and II) and segment B (helices II and III) from the X-ray structure during the CG MD simulation (replica 1). (b) Time series of the 100 ns all-atom CMD at 300 K in explicit solvent starting from the native structure. (c, d, e) Time series of the order parameters Φ for helices I, II, and III in the first replica of CG MD.

coordinates were reconstructed using CG-AA mapping. The heavy-atom RMSD of the reconstructed structures from the original ones were also examined. For main-chain mapping, the distribution of the RMSD had a very sharp peak at around the average value 0.55 Å, with a standard deviation of 0.01 Å, which can be considered as sufficiently small compared to thermal fluctuation. This is because the arrangement of the main-chain, with the exception of the termini, is almost determined by the geometrical condition of the C_{α} coordinates.⁵⁶ However, for side-chain mapping, the RMSD distribution had a broad peak averaging 1.97 Å, with a standard deviation of 0.21 Å. The coordinates of the generated side chains are determined as one of the optimal arrangements.⁵⁷ Since the alternative side-chain arrangement is

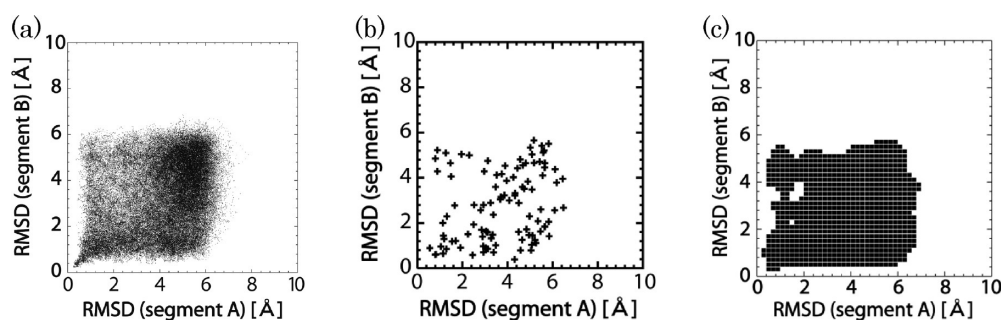


Figure 3. (a) Projections of the CG MD trajectories onto the subspace spanned by the C_{α} -RMSD of segment A and segment B and (b) those of the select 100 CG snapshots. (c) Overlapping regions among 100 distinct AA MD trajectories depicted by solid rectangles on the projected subspace. The overlap is counted if at least two distinct trajectories visit the rectangle. The size of each rectangle is $0.10 \text{ \AA} \times 0.10 \text{ \AA}$.

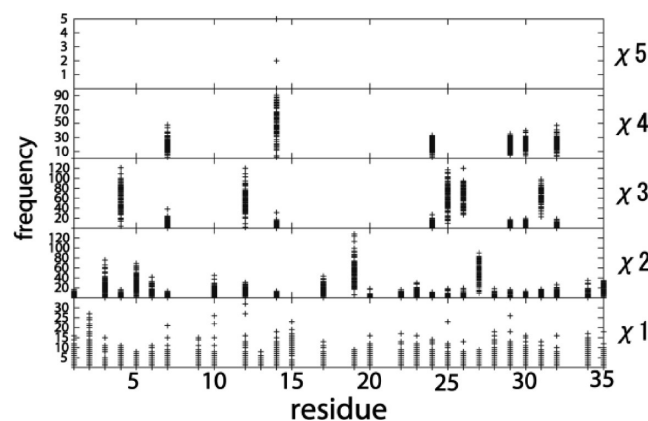


Figure 4. Transition frequencies of side-chain dihedral angles (χ_1 – χ_5) of each amino acid residue of HP35 in 100 1-ns AA MD simulations. Each symbol shows the transition frequency from one MD trajectory.

possible, we examined whether rotamer transitions occur frequently during AA MD, as was shown in the case of Trpzip2.²⁹ Figure 4 shows rotamer transition frequencies during a 1 ns MD of HP35. Highly exposed side chains showed very high transition frequencies. Even in the well-packed side chains, the transition frequencies were sufficient to take the other rotamer states. This result suggests that the calculated FEL has minimal dependence on the choice of initial side-chain arrangements.

After 100 independent AA MD simulations in the third stage, the FEL was calculated in the fourth stage using the WHAM. To calculate the FEL properly, AA MD trajectories should significantly overlap with a sufficient number of neighboring trajectories. In addition, the convergence of the calculated probability density should be examined. We calculated the number of overlaps between trajectories per trajectory, K . A pair of trajectories is regarded as overlapped if the C_{α} -RMSD between the average structures is smaller than 1.0 \AA . For each pair of overlapping trajectories, an all-to-all comparison is made among the snapshots, and the fraction of overlap Δ is estimated using the same C_{α} -RMSD criterion. One trajectory overlapped with $\bar{K} = 5.0 \pm 2.8$ trajectories, and $\Delta = 21.8 \pm 16.4\%$ of the snapshots overlapped in each pair of overlapping trajectories on average. We confirmed that all of the trajectories were not isolated and that all of the snapshots were connected in conformational space. To examine the convergence of \bar{K} and Δ values, 10 distinct series of the trajectories were prepared in random order. The value for the number of trajectories, $n = 30$, for example, indicates the quantity

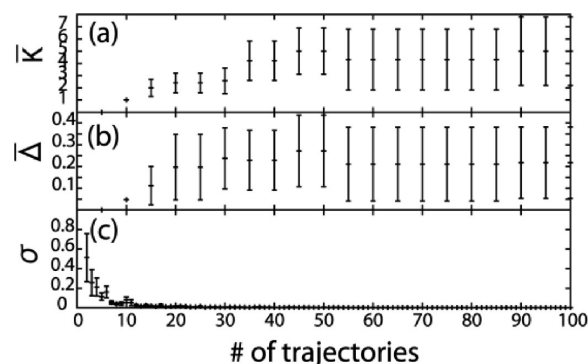


Figure 5. AA trajectory overlap and FEL convergence as a function of the number of trajectories considered, n . (a) The average number of overlapping trajectories per trajectory, \bar{K} . (b) The average fraction of overlapped snapshots between pairs of overlapping trajectories, Δ . (c) The convergence of the probability distribution projected onto the corresponding subspace, σ .

calculated with 30 trajectories and averaged over 10 distinct sets. The \bar{K} and Δ values were calculated as a function of n (Figures 5a,b). The \bar{K} value rapidly increased from 0 to 5, but the rate of increase became slower in the range $n \geq 40$ (Figure 5a). The values almost converged when 30 or more trajectories were considered (Figure 5b). We also examined the convergence of probability distributions projected onto a two-dimensional subspace spanned by the reaction coordinates. As shown in Figure 5c, σ almost converged at around $n = 30$, corresponding to the convergence of trajectory overlap. From these results, we concluded that the calculated FEL is well-converged with 100 trajectories.

To examine the convergence of the FEL versus the tertiary packing, we focused on the hydrophobic core formed as aromatic stacking of PHE6, PHE10, PHE17, and PHE35 shown in Figure 1. The fraction of the native contacts among the four phenylalanine residues (NC) was chosen as the reaction coordinate to describe the tertiary packing. To consider whether the 1 ns AA MD simulation is sufficient, one-dimensional FELs projected onto the NC calculated from the first and second halves of the all-atom 1 ns trajectories and their difference are shown in Figure 6. Since the difference is significantly smaller than $k_B T$, we judged that the 1 ns AA MD simulations were sufficient in length.

3.2. Major and Minor HP35 Folding Pathways. Figure 7a shows the folding FEL of HP35 obtained using the MSFEL method (1 ns \times 100 runs) projected onto the two-dimensional

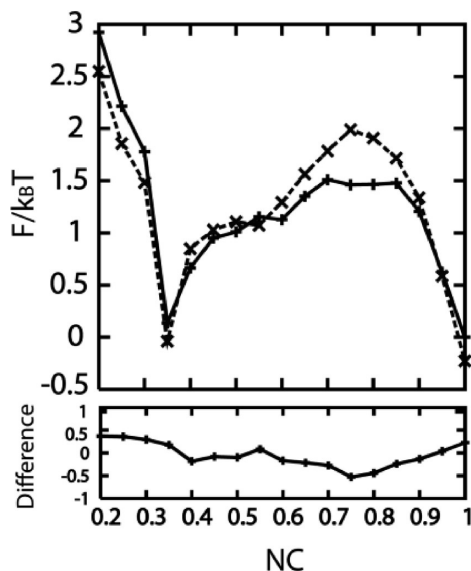


Figure 6. The one-dimensional free energy landscapes defined as the function of the fraction of the native contacts among hydrophobic core residues, PHE6, PHE10, PHE17, and PHE35. The solid and broken lines correspond to the free energy landscapes calculated from the first and second halves of the 1 ns trajectories, respectively. Hereafter, free energy value, F , is scaled by $k_B T$. Their difference is also shown below.

subspace spanned by the C_α -RMSD of two segments (R_A and R_B) from the crystal structure (PDB ID: 1YRF). In the folding FEL, four distinct states exist: denatured (D), major intermediate 1 (I1), minor intermediate 2 (I2), and native folded (N) (Figure 7a). To select the representative structures of these states, we first divided the subspace into four regions, D ($R_A > 2.0$ Å, $R_B > 2.0$ Å), I1 ($R_A > 2.0$ Å, $0 < R_B < 2.0$ Å), I2 ($0 < R_A < 2.0$ Å, $R_B > 2.0$ Å), and N ($0 < R_A < 2.0$ Å, $0 < R_B < 2.0$ Å), and calculated the weighted probability densities for grid points. The snapshot closest to the weighted average structure of the highest density grid in each region was selected as the representative structure.

To extract putative dynamic folding pathways, we first focused on the folding FEL (Figure 7a). As the first folding processes, the denatured protein folds into one of near intermediate states, I1 or I2. These two intermediate states can be distinguished from the denatured states by partial folding of segment A or B. Therefore, one-dimensional FEL projected onto the C_α -RMSD of each segment can describe the first folding processes, D→I1 (occurred in 1.0 Å $< R_A$) or D→I2 (occurred in 1.0 Å $< R_B$). The one-dimensional FEL on each first folding process was calculated as a double-well shape (Figure 9a,b). As the second folding processes, the intermediate structures fold into the native ones through partial folding of the remaining segment, I1→N (occurred in 1.0 Å $> R_B$) or I2→N (occurred in 1.0 Å $> R_A$). Thereby, the second folding process can be also described by the one-dimensional FEL projected onto the C_α -RMSD of each segment. The folding FEL on the second folding process also shows a double-well shape (Figure 9c,d). These pictures are the putative dynamical folding pathways extracted from the folding FEL shown in Figure 10a,b. As shown in Figure 9a,d, each state during D→I2→N was separated by relatively high free energy barriers compared to D→I1→N (Figure 9b,c). Therefore, D→I1→N is a more favorable route in the folding and defined as

the major folding pathway, whereas D→I2→N as the minor folding pathway.

In the major folding pathway (D→I1→N), segment B forms first (D→I1), and then the remaining region in segment A docks with segment B through I1 to reach N (I1→N). This can be interpreted as a two-step folding process. This major folding pathway agrees well with previous results calculated using REMD simulation (20 replicas \times 400 ns) with implicit solvent¹³ and multicanonical replica-exchange (MUCAREM) molecular dynamics (MD) simulations (total 2.28 μ s) with explicit solvent.⁴⁶

In this folding process, the formation of segment B precedes the hydrophobic residue contacts (Phe6, Phe10, and Phe17) that form the hydrophobic core necessary for stabilizing the native structure shown in Figure 1. In the minor folding pathway (D→I2→N) on the other hand, formation of segment A precedes formation of segment B. This minor folding pathway was not described in previous reports.^{12,13} Segment A in I2 is considered to be unstable because there is no hydrophobic side-chain stacking with segment B. As the energy barrier from I2 to N is relatively high ($\sim 6 k_B T$), the minor folding process from I2 to N would be expected to occur infrequently.

The folding FEL of HP35 indicates that the formation of I1 is the rate-limiting step in the process. Previous experimental work supports this hypothesis, suggesting that a well-conserved, solvent-exposed PLWK motif (residues 21–24) in segment B is critical for fast folding.⁵⁰ This site is considered to function as a structural gatekeeper in the HP35 folding process. The rigid Pro21 situated in the linker region between helices II and III plays a crucial role in restricting the movement of the two helices. The formation of this site initiates the folding of segment B.

To characterize the formation of the PLWK motif and segment B, segment B2 (residues 3–24) consisting of helix II and segment B3 (residues 21–33) consisting of helix III were examined for overlap of the PLWK motif (21–24) in the two segments. Characterization of overlap in this region enabled us to determine whether formation of the PLWK motif is correlated with the formation of segment B2 or segment B3. Figure 8 shows the FELs using the RMSD of segments B2 and B3 from the native structure as the reaction coordinates. In the major folding pathway, formation of segment B precedes formation of segment A (D→I1). Therefore, in the first stage of folding, the FEL can be calculated without considering the formation of segment A ($R_A > 2.0$ Å). Furthermore, the first stage of the folding FEL is divided into two types depending upon whether the motif forms or not. Figure 8a,b shows the FEL in the first stage of folding. The two folding FELs obviously show that segment B begins to form with the formation of the motif (Figure 8b) and that segment B does not form without formation of the motif (Figure 8a). These results support the experimentally derived hypothesis that the motif is the structural gatekeeper of HP35.⁵⁰ Figure 8b also shows that the formation of helix II is followed by the formation of helix III. Therefore, the following folding pathway is suggested: formation of the PLWK motif is triggered first, and then segment B is formed from C-terminal helix III to helix II.

We hypothesize that in the second folding stage (I1→N) formation of the hydrophobic core acts as a driving force for folding into the native state. To examine this possibility, fractions of native contacts between hydrophobic residues (Phe6, Phe10, and Phe17) and the C_α -RMSD of segment A were employed as reaction coordinates. The folding FEL of the second stage was calculated under the condition that segment B has already

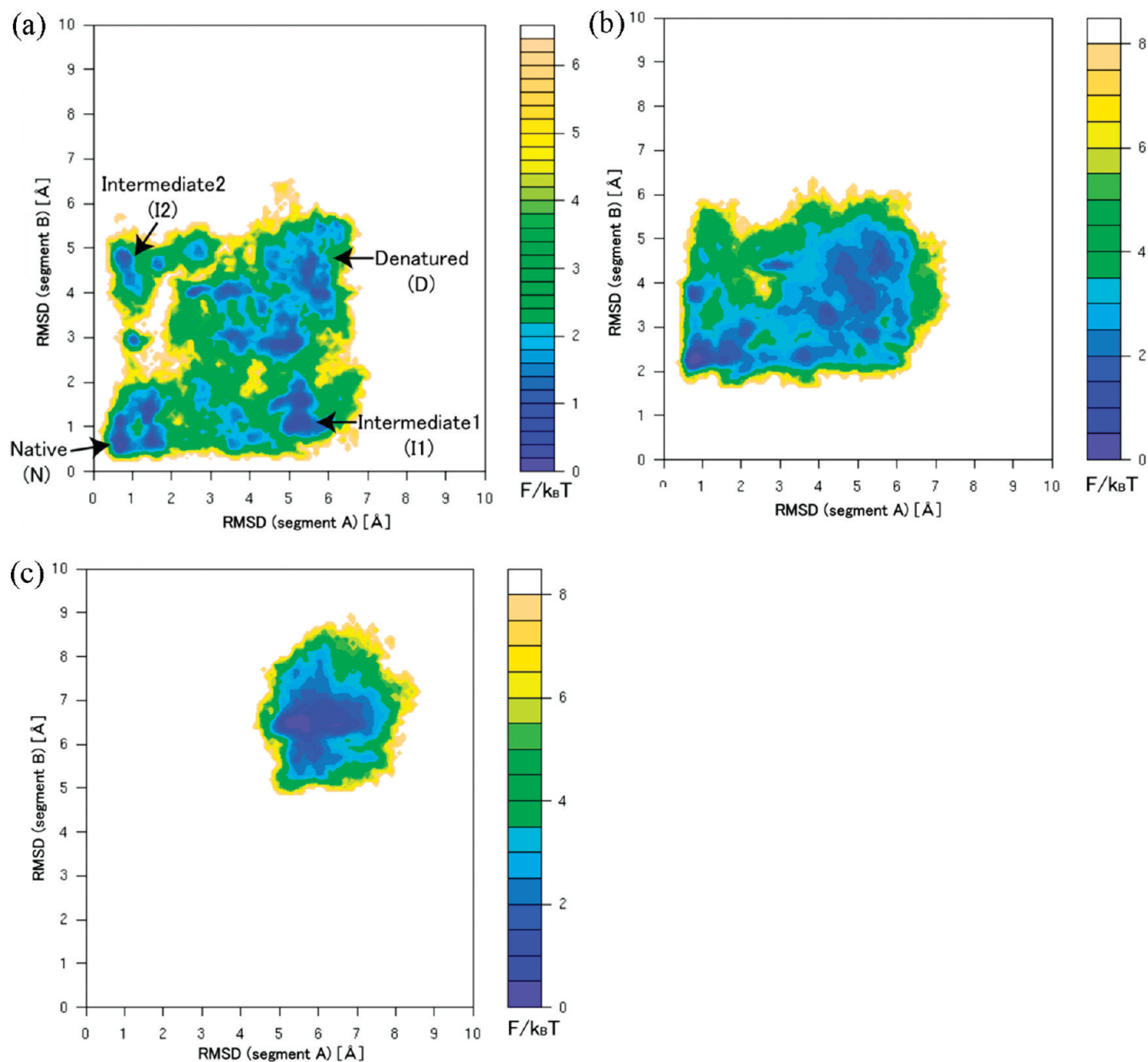


Figure 7. (a) The folding FEL of HP35 projected onto the subspace spanned by the C_{α} -RMSD of segment A and segment B calculated using the MSFEL method. (b) Folding FEL of the P21A and (c) F6A/F10A/F17A mutants.

formed ($R_B < 2.0$ Å). Figure 8c shows the folding FEL in the second stage of folding and indicates that the formation of segment A is dependent upon formation of the hydrophobic native contacts. The calculated free energy difference during the folding process ($\sim 4 k_B T$) was slightly smaller than the experimentally derived value ($\sim 5 k_B T$).⁴⁰ This is a better agreement than previous work with implicit solvent.¹³

In the minor folding pathway, the order of segment formation is reversed; the formation of segment A ($D \rightarrow I2$) precedes the formation of segment B ($D \rightarrow I1$). The trigger for the minor pathway is contact of the hydrophobic residues to form the hydrophobic core, not the formation of the motif as in the major pathway. Figure 8d shows the FEL of the first folding stage in the minor folding pathway when segment B was not formed ($R_B > 2.0$ Å). After the formation of segment A ($R_A < 2.0$ Å), segment B

is formed through the formation of the motif in the same order (helix III \rightarrow helix II) as shown in Figure 8e,f. Graphical summary representations of the major and minor folding pathways are shown in Figure 10a,b. This minor folding pathway is consistent with the recent result of triplet–triplet-energy transfer (TTET) experiment.⁶⁴ The high-free-energy intermediate found to be accessible from the N state is expected to correspond to I2 in the MSFEL. The experimental activation barrier between D and I2 was reported to be $\sim 7 k_B T$,⁶⁴ which is comparable to our calculation $\sim 6 k_B T$. The slight underestimation may be due to the choice of the force field. It has been reported that the AMBER ff03 force field had higher helical stability than in experiments. Folding enthalpies were also less than half of the experimental value.⁶⁵ Furthermore, it has been pointed that the ff03 favors a helical unfolded state and a diffusion-collision-type folding mechanism.

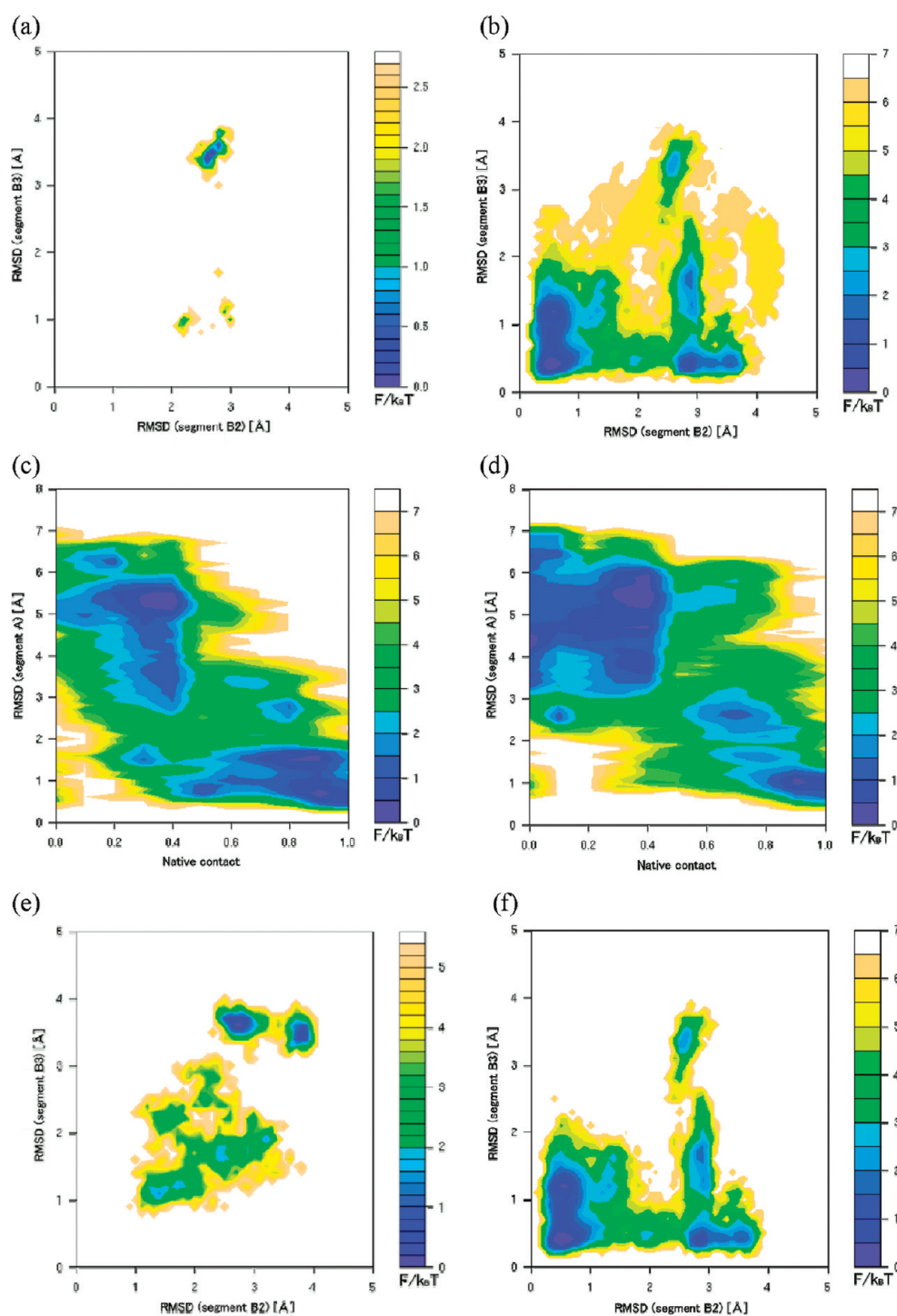


Figure 8. (a, b) The folding FEL of HP35 in the first stage of the major pathway ($D \rightarrow I1$, $R_A > 2.0 \text{ \AA}$) projected onto the two-dimensional subspace spanned by the C_α -RMSD of segments B2 and B3, (a) without considering the formation of the PLWK motif and (b) considering formation of the PLWK motif. (c) The folding FEL of HP35 in the second stage of the major pathway ($I1 \rightarrow N$, $R_A < 2.0 \text{ \AA}$) projected onto the two-dimensional subspace spanned by fractions of the native contact between the hydrophobic residues (Phe6, Phe10, and Phe17) and the C_α -RMSD of segment A. (d) The folding FEL of HP35 in the first stage of the minor pathway ($D \rightarrow I2$, $R_B > 2.0 \text{ \AA}$) projected onto the same reaction coordinates as for c. The folding FEL of the second stage in the minor pathway ($D \rightarrow I2$, $R_A < 2.0 \text{ \AA}$) (e) when the motif was not formed and (f) when the motif was formed.

Individual helices were rather stable in isolation and dock together to form the folded state. Actually, this interpretation agrees with the folding pathways shown as Figure 10a,b; each segment docks together into the native conformation when each helix is almost folded. Therefore, the stability of helical structures

with the ff03 force field may lead the underestimation of the D to I2 free energy difference.

The folding pathways observed in this work are also compared to the result of all-atom unbiased 100 μ s MD simulations⁶⁶ in which the same force field (AMBER ff03) was employed.

The folding free energy barrier height along the 1D major pathway at 300 K derived from Figure 9b,c corresponds to ~ 1.2 kcal/mol ($\sim 2.0 k_B T$), which is 6.0 times higher than the value at 390 K (~ 0.2 kcal/mol).⁶⁶ This difference will be originated from the temperature difference because the frequent

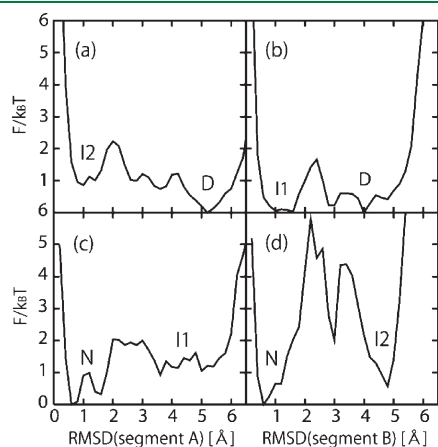


Figure 9. (a,b) The one-dimensional free energy landscapes projected onto the C_{α} -RMSD of segment A and B on the first folding processes of the minor and major folding pathways and (c,d) those of the second folding processes of the major and minor folding pathways.

observation of the folding–unfolding transitions at 390 K should be caused by the reduction of the free energy barrier. The order of helix formation on the major folding pathway derived from Figure 8 showed good agreement with that of the all-atom MD simulation in the following points: (1) helix 1 is relatively unstable in the unfolded state; (2) helices 2 and 3 form during the early stage of the folding process; and (3) helix 1 is nearly always the last to form.

3.3. Mutation in the PLWK Motif. The importance of Pro21 in the PLWK motif has been shown by the point mutation experiments.^{50,51} The point mutation P21A had a dramatic effect on the folding of HP35. It has also been suggested that Pro21 may be responsible for critical interactions for folding into the native structure. To examine the effect of this mutation in the PLWK motif from a point of view of FEL, we constructed a mutant (P21A) and calculated its folding FEL (Figure 7b). The formation of segment B was clearly suppressed in this mutant, indicating that the rigid Pro21 in the linker region between helices II and III plays an important role in stabilizing segment B. This mutation only suppressed the formation of segment B; therefore, Pro21 is essential for controlling the major folding pathway through its influence on the formation of the PLWK motif. This result was in good agreement with previous NMR and CD experimental and computational results of the point mutation.⁵⁰

3.4. Mutation in the Hydrophobic Core. To address the effect of mutations in the hydrophobic core shown in Figure 1, we constructed a triple mutant, F6A/F10A/F17A. Figure 7c

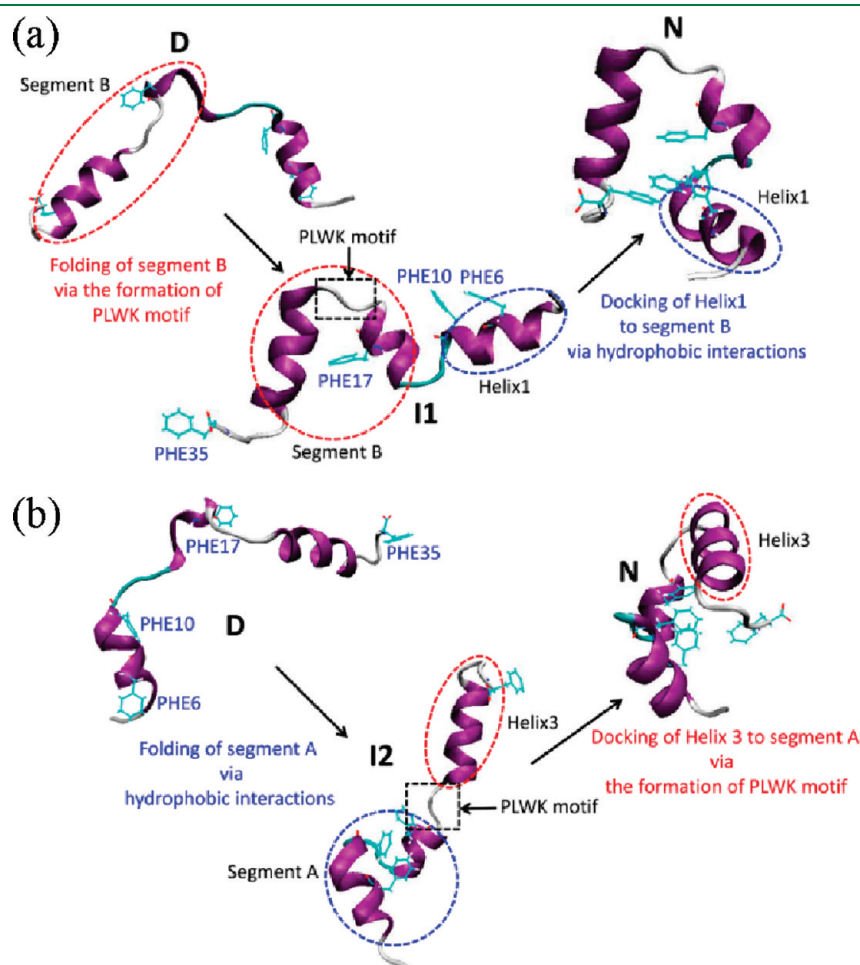


Figure 10. Schematic representations of (a) the major and (b) minor folding pathways. Figure created using VMD.⁶⁷

shows the folding FEL of the mutant. The mutations are expected to suppress the formation of segment A as the formation of the hydrophobic core is the driving force behind the formation of segment A in the minor folding pathway. The folding FEL shown in Figure 7c indicated that there was suppression of the formation of both segments A and B. These observations agree well with the previously reported results of mutagenesis experiments,⁴² suggesting that these three phenylalanine residues play crucial roles in stabilizing the native structure and in the folding of HP35.

4. CONCLUSION

In the present work, we studied the folding process of the fast-folding mini-protein HP35 by investigating the folding FEL in explicit solvent using the MSFEL method. A previously unreported minor folding pathway in the order of D→I2→N was identified in this work in addition to the major folding pathway, D→I1→N, described previously.^{12,13} In the minor pathway, the driving force behind folding is the formation of the hydrophobic core (residues F6, F10, and F17) located at the center of the native structure, while the formation of the PLWK motif (residues P21–K24) is considered to be the trigger in the major pathway. Mutations in the PLWK motif (P21A) and hydrophobic core showed that these residues play important roles in the folding of HP35. The P21A mutation partially suppressed folding, especially the formation of helices II and III, while mutations in the hydrophobic core completely prevented the overall folding of HP35.

AUTHOR INFORMATION

Corresponding Author

*Phone/Fax: +81-3-5841-2297. E-mail: kitao@iam.u-tokyo.ac.jp.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

This work was supported by the Next Generation Super Computing Project, Nanoscience Program, the Strategic Programs for Innovative Research (SPIRE), and Computational Material Science Initiative (CMSI), Grants-in-Aid for Science Research (B), and Grants-in-Aid for Science Research in Priority Areas from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan to A.K. The computations were performed in part using supercomputers at the Research Center for Computational Science, Okazaki Research Facilities, National Institute of Natural Science.

REFERENCES

- (1) Bevivino, A. E.; Loll, P. J. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 11955–11960.
- (2) Cooper, J. K.; Schilling, G.; Peters, M. F.; Herring, W. J.; Sharp, A. H.; Kaminsky, Z.; Masone, J.; Khan, F. A.; Delanoy, M.; Borchelt, D. R.; Dawson, V. L.; Dawson, T. M.; Ross, C. A. *Hum. Mol. Genet.* **1998**, *7*, 783–790.
- (3) Georgalis, Y.; Starikov, E. B.; Hollenbach, B.; Lurz, R.; Scherzinger, E.; Saenger, W.; Lehrach, H.; Wanker, E. E. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 6118–6121.
- (4) Scherzinger, E.; Lurz, R.; Turmaine, M.; Mangiarini, L.; Hollenbach, B.; Hasenbank, R.; Bates, G. P.; Davies, S. W.; Lehrach, H.; Wanker, E. E. *Cell* **1997**, *90*, 549–558.
- (5) Singer, S. J.; Dewji, N. N. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 1546–1550.

- (6) Satoh, D.; Shimizu, K.; Nakamura, S.; Terada, T. *FEBS Lett.* **2006**, *580*, 3422–3426.
- (7) Suenaga, A.; Narumi, T.; Futatsugi, N.; Yanai, R.; Ohno, Y.; Okimoto, N.; Taiji, M. *Chem.—Asian J.* **2007**, *2*, 591–598.
- (8) Juraszek, J.; Bolhuis, P. G. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 15859–15864.
- (9) Zhou, R. H. *Proteins: Struct., Funct., Genet.* **2003**, *53*, 148–161.
- (10) Duan, Y.; Kollman, P. A. *Science* **1998**, *282*, 740–744.
- (11) Freddolino, P. L.; Schulten, K. *Biophys. J.* **2009**, *97*, 2338–2347.
- (12) Lei, H. X.; Duan, Y. *J. Mol. Biol.* **2007**, *370*, 196–206.
- (13) Lei, H. X.; Wu, C.; Liu, H. G.; Duan, Y. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 4925–4930.
- (14) Nakajima, N.; Nakamura, H.; Kidera, A. *J. Phys. Chem. B* **1997**, *101*, 817–824.
- (15) Hansmann, U. H. E.; Okamoto, Y.; Eisenmenger, F. *Chem. Phys. Lett.* **1996**, *259*, 321–330.
- (16) Hukushima, K.; Nemoto, K. *J. Phys. Soc. Jpn.* **1996**, *65*, 1604–1608.
- (17) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.
- (18) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087–1092.
- (19) Sugita, Y.; Kitao, A.; Okamoto, Y. *J. Chem. Phys.* **2000**, *113*, 6042–6051.
- (20) Bolhuis, P. G.; Juraszek, J. *Biophys. J.* **2010**, *98*, 646–656.
- (21) Laio, A.; Gervasio, F. L. *Rep. Prog. Phys.* **2008**, *71*.
- (22) Wang, F. G.; Landau, D. P. *Phys. Rev. Lett.* **2001**, *86*, 2050–2053.
- (23) Dellago, C.; Grunwald, M. *J. Chem. Phys.* **2007**, *127*.
- (24) Garcia, A. E.; Onuchic, J. N. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 13898–13903.
- (25) Gnanakaran, S.; Nymeyer, H.; Portman, J.; Sanbonmatsu, K. Y.; Garcia, A. E. *Curr. Opin. Struct. Biol.* **2003**, *13*, 168–174.
- (26) Jang, S.; Kim, E.; Pak, Y. *J. Chem. Phys.* **2008**, *128*, 105102.
- (27) Zhang, J.; Qin, M.; Wang, W. *Proteins: Struct., Funct., Bioinf.* **2006**, *62*, 672–685.
- (28) Kitao, A. *J. Chem. Phys.* **2011**, *135*, 045101.
- (29) Harada, R.; Kitao, A. *Chem. Phys. Lett.* **2011**, *503*, 145–152.
- (30) Torrie, G. M.; Valleau, J. P. *Chem. Phys. Lett.* **1974**, *28*, 578–581.
- (31) Torrie, G. M.; Valleau, J. P. *J. Comput. Phys.* **1977**, *23*, 187–199.
- (32) Ferrenberg, A. M.; Swendsen, R. H. *Phys. Rev. Lett.* **1989**, *63*, 1195–1198.
- (33) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011–1021.
- (34) Souaille, M.; Roux, B. *Comput. Phys. Commun.* **2001**, *135*, 40–57.
- (35) Harada, R.; Kitao, A. *J. Phys. Chem. B* **2011**, *115*, 8806–8812.
- (36) Tang, Y. F.; Grey, M. J.; McKnight, J.; Palmer, A. G.; Raleigh, D. P. *J. Mol. Biol.* **2006**, *355*, 1066–1077.
- (37) Vardar, D.; Chishti, A. H.; Frank, B. S.; Luna, E. J.; Noegel, A. A.; Oh, S. W.; Schleicher, M.; McKnight, C. J. *Cell Motil. Cytoskeleton* **2002**, *52*, 9–21.
- (38) Chiu, T. K.; Kubelka, J.; Herbst-Irmer, R.; Eaton, W. A.; Hofrichter, J.; Davies, D. R. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 7517–7522.
- (39) McKnight, C. J.; Matsudaira, P. T.; Kim, P. S. *Nat. Struct. Biol.* **1997**, *4*, 180–184.
- (40) Kubelka, J.; Eaton, W. A.; Hofrichter, J. *J. Mol. Biol.* **2003**, *329*, 625–630.
- (41) Kubelka, J.; Chiu, T. K.; Davies, D. R.; Eaton, W. A.; Hofrichter, J. *J. Mol. Biol.* **2006**, *359*, 546–553.
- (42) Frank, B. S.; Vardar, D.; Buckley, D. A.; McKnight, C. J. *Protein Sci.* **2002**, *11*, 680–687.
- (43) Lei, H.; Deng, X.; Wang, Z.; Duan, Y. *J. Chem. Phys.* **2008**, *129*, 155104.
- (44) Piana, S.; Laio, A.; Marinelli, F.; Van Troys, M.; Bourry, D.; Ampe, C.; Martins, J. C. *J. Mol. Biol.* **2008**, *375*, 460–470.
- (45) Raleigh, D. P.; Wickstrom, L.; Okur, A.; Song, K.; Hornak, V.; Simmerling, C. L. *J. Mol. Biol.* **2006**, *360*, 1094–1107.
- (46) Yoda, T.; Sugita, Y.; Okamoto, Y. *Biophys. J.* **2010**, *99*, 1637–1644.

- (47) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y. B.; Wriggers, W. *Science* **2010**, *330*, 341–346.
- (48) Schulten, K.; Freddolino, P. L. *Biophys. J.* **2009**, *97*, 2338–2347.
- (49) Kollman, P. A.; Duan, Y. *Science* **1998**, *282*, 740–744.
- (50) Vermeulen, W.; Van Troys, M.; Bourry, D.; Dewitte, D.; Rossenu, S.; Goethals, M.; Borremans, F. A. M.; Vandekerckhove, J.; Martins, J. C.; Ampe, C. *J. Mol. Biol.* **2006**, *359*, 1277–1292.
- (51) Raleigh, D. P.; Xiao, S. F. *J. Mol. Biol.* **2010**, *401*, 274–285.
- (52) Brooks, C. L., 3rd *Curr. Opin. Struct. Biol.* **1998**, *8*, 222–226.
- (53) Go, N. *Annu. Rev. Biophys. Bioeng.* **1983**, *12*, 183–210.
- (54) Mirny, L.; Shakhnovich, E. *Annu. Rev. Biophys. Biomol. Struct.* **2001**, *30*, 361–396.
- (55) Martyna, G. J.; Klein, M. L.; Tuckerman, M. J. *Chem. Phys.* **1992**, *97*, 2635–2643.
- (56) Gront, D.; Kmiecik, S.; Kolinski, A. *J. Comput. Chem.* **2007**, *28*, 1593–1597.
- (57) Canutescu, A. A.; Shelenkov, A. A.; Dunbrack, R. L. *Protein Sci.* **2003**, *12*, 2001–2014.
- (58) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (59) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani, F.; Wu, X.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Matthews, D. H.; Schafmeister, C.; Ross, W. S.; Kollman, P. A. *AMBER 9*; University of California, San Francisco, 2006.
- (60) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G. M.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J. M.; Kollman, P. *J. Comput. Chem.* **2003**, *24*, 1999–2012.
- (61) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.
- (62) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593.
- (63) Frishman, D.; Argos, P. *Proteins: Struct., Funct., Genet.* **1995**, *23*, 566–579.
- (64) Kiefhaber, T.; Reiner, A.; Henklein, P. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 4955–4960.
- (65) Shaw, D. E.; Piana, S.; Lindorff-Larsen, K. *Biophys. J.* **2011**, *100*, L47–L49.
- (66) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. *Biophys. J.* **2011**, *100*, L47–L49.
- (67) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33–38.

Parameterization of PACE Force Field for Membrane Environment and Simulation of Helical Peptides and Helix–Helix Association

Cheuk-Kin Wan,[†] Wei Han,[†] and Yun-Dong Wu^{*,†,‡,§}

[†]Department of Chemistry, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, China

[‡]School of Chemical Biology and Biotechnology, Laboratory of Chemical Genomics, Peking University Shenzhen Graduate School, Shenzhen, 518055, China

[§]College of Chemistry, Peking University, Beijing, 100871, China

 Supporting Information

ABSTRACT: The recently developed PACE force field was further parametrized so that it can be applied to the studies of membrane systems. Parameters for the interactions between united-atom protein particles and lipid hydrophobic tails were developed by reproducing the solvation free energies of small organic molecules in hexadecane. Interactions between protein particles and lipid heads were parametrized by fitting the potential of mean force of the corresponding all-atom simulation. The force field was applied to the study of five helical peptides in membrane environments. The calculated tilt angles of WALP and GWALP and their mutations are in good agreement with experimental data. The association of two glycoporphin A (GpA) helices was simulated for 6 μ s. Root-mean-square-deviation of the simulated dimer from the nuclear magnetic resonance structure was found to be 0.272 nm, better than all results obtained so far. These findings demonstrate the high accuracy and applicability of the PACE force field in studying membrane proteins.

INTRODUCTION

Membrane proteins are important for many cellular processes, including transport activities, signal transductions, and receptor functions.¹ Understanding structures, dynamics, and functions of membrane proteins is challenging.^{2–4} For instance, although the first crystal structure of membrane proteins was obtained in 1985, only about 300 unique crystal structures have been determined in the 26 years since (see <http://blanco.biomol.uci.edu/mpstruc/listAll/list> for details).⁵ Molecular dynamics (MD) simulation is a complementary tool to reveal the structural and dynamic details of proteins.^{6–8} It can also help with the interpretation and evaluation of experimental findings.⁹ All-atom force fields are commonly used but limited to relatively short simulation lengths. The folding and dynamics of membrane proteins are in the time scale of microseconds or even longer. Although continuous trajectories were achieved in 10 μ s and 1 ms simulations by Schulten et al.¹⁰ and Shaw et al.¹¹ for the folding of soluble proteins, they relied on the use of supercomputers and specially designed machines not available to most people.

Coarse-grained (CG) force fields have been actively developed to reduce computational cost for more efficient long time simulations. Klein et al. grouped three heavy atoms and their associated hydrogen atoms into one CG particle. This model was used to study the membrane insertion activity of antimicrobial polymers.^{12,13} Sperotto et al. developed a similar CG model for the simulation of lipid-mediated protein–protein interactions.^{14,15} Apart from coarse-graining three heavy atoms into one CG particle, a four-to-one mapping has also been used to simplify lipid molecules to investigate vesicle fission and fusion.^{16,17} Voth et al. developed a solvent-free lipid bilayer model using a multiscale coarse-graining approach to enhance computational efficiency.¹⁸ Another four-to-one CG model—the

MARTINI force field—that has found extensive applications was developed by Marrink and co-workers.^{19–21} It has been applied to biomolecular simulations, such as membrane fusion and the spontaneous gating of channel.^{22,23} This model has been extended to include different types of particles. The first extended protein force field was built by Schulten et al. based on the first version of MARTINI lipid model to study lipoproteins.²⁴ Sansom et al. developed another protein model for membrane protein simulations.²⁵ Tarek et al. parametrized cyclic peptides to describe the self-assembly of [Trp-Leu]₄.²⁶

The common feature of these CG models is the similar degree of coarse graining of proteins and solvent molecules. Although these CG approaches can speed up MD simulations by 2–3 orders of magnitude, they may not be accurate enough due to oversimplification and the loss of atomistic details. Details of solvent molecules are thought to be less significant, so CG or even implicit solvent models are generally employed in studying proteins. Atomistic details of protein molecules are more important in maintaining and predicting protein structures. Coarse graining of proteins is therefore less desirable. To solve this problem, a finer protein model could be incorporated into a CG solvent model to build a hybrid-resolution protein force field. Voth et al. built a mixed all-atom and CG model for gramicidin A ion channel simulation.²⁷ We proposed a hybrid-resolution force field (PACE) to couple a united-atom (UA) protein model with the MARTINI CG environment. A similar approach was recently proposed by Marrink et al. which was implemented differently.²⁸ In our previous work, we optimized the PACE force field and coupled it with the CG water model.^{29–32} We showed that the

Received: June 21, 2011

Published: November 08, 2011

force field could reproduce the statistical backbone and side chain potentials of all amino acids accurately.³³ We demonstrated that not only could it maintain the stability of the native structures of proteins with medium size (50–150 amino acids), it could also fold several α -helical, β -sheet, and mixed helical/coil peptides from first principles.

Encouraged by these results, we continue to extend the PACE force field to cover the CG membrane environment. In this paper, we report our attempt to incorporate MARTINI's lipid model into our PACE force field.²⁰ The parametrizations involved include interactions between protein UA particles and lipid CG hydrophobic tail groups and protein UA particles and lipid CG head groups. To evaluate the quality of the modified PACE force field, we applied it to the simulations of WALP, a designed helical peptide that has been widely studied both experimentally and theoretically,^{34–43} and four other helical peptides.^{44,45} We pay particular attention to the tilt angle of these peptides in two membrane environments. We also studied the dimerization of glycyphorin helix A (GpA) in membrane, which is currently a topic of intense interest.^{46–69} Our simulations reproduce related experimental observations quite well, which implies that the extended PACE force field may have potential applications in the study of membrane proteins with the atomistic details.

MODEL AND METHODS

PACE Protein Model. In the PACE force field, proteins are represented at a UA level and embedded in a CG environment.³¹ Figure 1a shows the schematic representation of our model. The water and lipid models are adopted from the MARTINI model.²⁰ Around four heavy atoms are represented by a CG particle. The total energy of the MARTINI model is expressed in eqs 1–3:

$$E_{CG\text{-total}} = E_{CG\text{-bonded}} + E_{CG\text{-nonbonded}} \quad (1)$$

$$E_{CG\text{-bonded}} = E_{CG\text{-bond}} + E_{CG\text{-angle}} \quad (2)$$

$$E_{CG\text{-nonbonded}} = E_{CG\text{-vdW}} + E_{CG\text{-electrostatic}} \quad (3)$$

Bond lengths and angles are modeled by harmonic potentials $E_{CG\text{-bond}}$ and $E_{CG\text{-angle}}$, respectively. Lennard-Jones potential energy function (eq 4) is used to describe van der Waals interactions ($E_{CG\text{-vdW}}$) between CG particles, while ϵ_{ij} indicates the strength of the interaction, and δ_{ij} indicates the distance between two interacting groups with zero interaction energy.

$$E_{CG\text{-vdW}} = \sum_{i \neq j} 4\epsilon_{ij} \left(\frac{\delta_{ij}^{12}}{r^{12}} - \frac{\delta_{ij}^6}{r^6} \right) \quad (4)$$

Electrostatic interactions ($E_{CG\text{-electrostatic}}$) between charged groups are modeled by Coulombic potential energy function (eq 5), and q_i and q_j are charges of the charged groups. Relative dielectric constant ϵ_r is set to be 15 for explicit screening.

$$E_{CG\text{-electrostatic}} = \sum_{i \neq j} \frac{q_i q_j}{4\pi\epsilon_0\epsilon_r r} \quad (5)$$

eqs 1–5 are used for CG–CG interactions. All parameters are adopted from the MARTINI model and can be found in ref 20.

The protein model of PACE is UA based in that hydrogen atoms are implicitly incorporated into the attached heavy atoms.

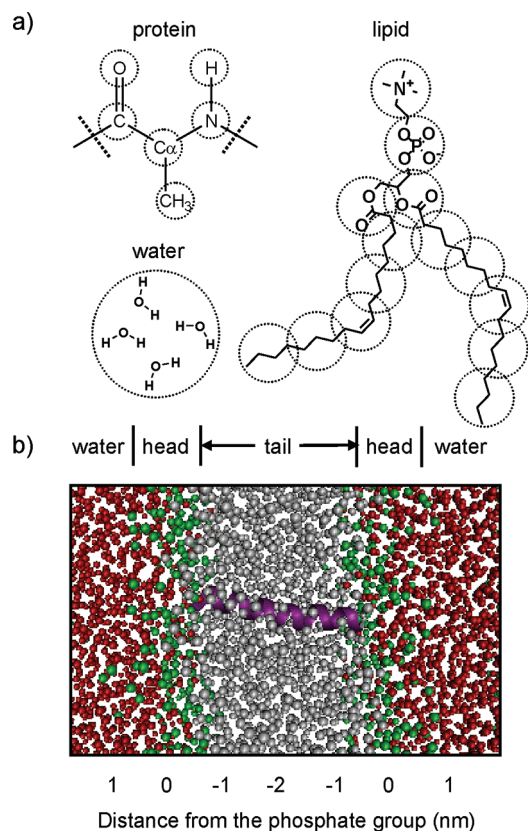


Figure 1. (a) Schematic representation of the PACE model. Protein particles and the environment (lipid bilayer and water) are represented by the UA and CG models respectively. (b) Snapshot of a peptide in the DOPC bilayer. The UA protein is shown in purple. The CG lipid tail, lipid head, and water particles are shown in gray, green, and red, respectively.

In order to have a good account of the hydrogen bonds, we represent explicitly the hydrogens that are attached to nitrogen for the backbone amide and side chains of Asn, Gln, Trp, and His. Detailed schematic representations of 20 amino acids can be found in ref 31. For UA–UA interactions, we used similar potential energy functions to describe bonded and nonbonded interactions. The total energy of our protein model is expressed in eqs 6 and 7:

$$E_{UA\text{-total}} = E_{UA\text{-bonded}} + E_{UA\text{-nonbonded}} + E_{UA\text{-HB}} \quad (6)$$

$$E_{UA\text{-bonded}} = E_{UA\text{-angle}} + E_{UA\text{-improper}} + E_{UA\text{-torsion}} + E_{UA\text{-14pair}} \quad (7)$$

All bond lengths are constrained by the LINCS algorithm.⁷⁰ Bond angles are constrained by a harmonic potential $E_{UA\text{-angle}}$ at their equilibrium. Planar geometries and chiral centers of molecules are maintained by $E_{UA\text{-improper}}$, $E_{UA\text{-torsion}}$ and $E_{UA\text{-14pair}}$ describe the potential energy of the dihedral angle of a rotatable bond. All optimized parameters can be found in ref 32.

None of the protein particles of PACE carry charges. A single potential energy function, eq 4 is used to describe nonbonded interactions ($E_{UA\text{-nonbonded}}$) between UA particles. The parameters for protein–water interactions were optimized by fitting the hydration free energies of 35 organic compounds in a previous study. The average deviation from experimental data

was 1.1 kJ/mol.³² The interactions between protein particles were parametrized by reproducing the experimental density and self-solvation free energy of eight pure organic liquids. The average errors for density and self-solvation free energy were 3.2% and 0.7 kJ/mol, respectively. Side chain–side chain and side chain–backbone interaction potentials were parametrized to fit the potential of mean force of corresponding all-atom simulations using the OPLS-AA force field. The results and the optimized parameters can be found in ref 31.

$E_{\text{UA-HB}}$ is used to control the strength of backbone–backbone H-bond interactions and is given by eq 8. Parameters are optimized by reproducing the experimental α -helical and β -sheet content of AK17 ((AAKAA)₃GY) and GB1m2 (GEWTYN-PATGKFTVTE) peptides.³¹ Note that we use a slightly different potential to describe side chain H-bonds, as the H-bond potentials for backbones can only handle hydrogen donors containing only one hydrogen atom. These parameters are obtained by fitting the all-atom potential of mean forces (PMFs). The schematic representation of the different H-bond types is shown in Figure S1, Supporting Information.

$$E_{\text{UA-HB}} = \sum_{|i-j|>2} \left[4\epsilon_{\text{attr}} \left(\frac{\delta_{\text{O}_i-\text{NH}_j}^{12}}{r_{\text{O}_i-\text{NH}_j}^{12}} - \frac{\delta_{\text{O}_i-\text{NH}_j}^6}{r_{\text{O}_i-\text{NH}_j}^6} \right) + 4\epsilon_{\text{rep}} \frac{\delta_{\text{O}_i-\text{C}_{\alpha j}}^{12}}{r_{\text{O}_i-\text{C}_{\alpha j}}^{12}} + 4\epsilon_{\text{rep}} \frac{\delta_{\text{O}_i-\text{C}_j-1}^{12}}{r_{\text{O}_i-\text{C}_j-1}^{12}} + 4\epsilon_{\text{rep}} \frac{\delta_{\text{C}_i-\text{NH}_j}^{12}}{r_{\text{C}_i-\text{NH}_j}^{12}} \right] \quad (8)$$

To incorporate the UA protein model into the CG lipid model, parametrizations are carried out for nonbonded interactions between protein–lipid tail and protein–lipid head groups. eq 4 is used to describe UA–CG interactions. Therefore, ϵ_{ij} and δ_{ij} are carefully optimized in this work.

As most of the computation time is spent on simulations of the solvent and the lipid bilayer, simplification of the solvent and lipid model can reduce the number of degrees of freedom and greatly enhance the computation efficiency. Note that the effective times of the MARTINI model are normally interpreted as four times the simulation times.²¹ Our protein model is UA based, and the effective times can only be interpreted after further studies. In this work, all reported simulation lengths are actual simulation times.

Simulation Parameters. All of the simulations were performed with the GROMACS software package, version 3.3.1.⁷¹ The van der Waals interactions were shifted to zero between distances of 0.9 and 1.2 nm. The electrostatic forces were shifted from 0.0 to 1.2 nm with a dielectric constant of 15 for explicit screening. The neighbor list was updated every 10 steps with a searching cutoff distance of 1.4 nm. The temperature of each group (protein, lipid and water) was kept constant using the Berendsen thermostat⁷² with a time constant of 0.1 ps. The pressure on the z -axis and that on the xy plane were each coupled using the Berendsen barostat⁷² with a time constant of 0.5 ps and a compressibility of $4.5 \times 10^{-5} \text{ bar}^{-1}$. In all simulations, the normal of the membrane was aligned along the z -axis. All of the systems were kept at 1 atm. As explicit hydrogen atoms are represented by dummy atoms, a larger time step of 6 fs can be used, which has been shown to satisfy energy conservation during simulations.⁷³ In the preparation stage, 5000 steps of steep descent optimization were performed, followed by a 1 ns

pre-equilibrium simulation at 300 K and 1 atm. The generated coordinates and velocities were used for simulation runs.

Solvation Free Energy Calculations. The solvation free energy is the change in free energy of a molecule from the ideal gas state to a state where the molecule is immersed in the solvent. It is calculated by introducing a coupling parameter (λ) with interaction potentials between the solute and the solvent using the free energy perturbation method. In the simulations, as λ gradually changed from zero to unity, the solute–solvent interactions were gradually turned off. The solvation free energy was then calculated as follows:⁷⁴

$$\Delta G = G_1 - G_0 = \int_{\lambda=0}^{\lambda=1} d\lambda \left\langle \frac{\partial U(\lambda)}{\partial \lambda} \right\rangle_{\lambda} \quad (9)$$

The solute molecule was put into a dodecahedron box containing 400 solvent molecules. For each solvation free energy calculation, 24 intermediate stages were applied ($\lambda = 0.0, 0.06, 0.15, 0.25, 0.35, 0.4, 0.45, 0.5, 0.53, 0.56, 0.59, 0.62, 0.65, 0.68, 0.71, 0.73, 0.75, 0.78, 0.81, 0.84, 0.88, 0.92, 0.96, 1.0$). A 6 ns simulation was performed for each intermediate stage giving a total of 144 ns of simulations. The temperature was kept at 300 K. The reproducibility of the calculations was tested. The average deviation of the solvation free energy calculations in several runs is about 1 kJ/mol.

PMF Calculations. For each PMF calculation, a small molecule was put in the center of the dioleoylphosphatidylcholine (DOPC) bilayer containing 70 lipid molecules and 1100 CG water molecules. PMFs were calculated using the free energy perturbation method. The position of the center of mass of a molecule was fixed by a potential $k/2(x - x_0)^2$. The x_0 can be perturbed by introducing a λ . As λ varied from zero to unity, the solute was moved from the center of the lipid bilayer to the aqueous phase in a straight line along the bilayer normal. The free energy in the aqueous phase ($\lambda = 1$) was chosen as the reference state. The free energy difference between any intermediate point (λ_i) and the starting point ($\lambda = 0$) can be estimated by accumulating $dU/d\lambda$ until $\lambda = \lambda_i$. In each simulation, the free energies of 100 intermediate states were chosen for PMF plots. Each simulation was carried out for 400 ns. For each PMF calculation, 10 perturbation simulations were performed to generate an average PMF curve, leading to a total of 4 μs of simulations. All simulations were kept at 300 K throughout the entire duration.

Transmembrane Peptide Simulations. Transmembrane peptides were inserted into the pre-equilibrated lipid bilayer, dilauroylphosphatidylcholine (DLPC), dipalmitoylphosphatidylcholine (DPPC), or DOPC, of 128 lipids with 1500 CG water molecules. The starting structures of the peptides were fully α -helical and in orientation of about 0° tilt relative to the bilayer normal. To have the peptides vertically inserted into the lipid bilayer for MD simulations, different setup methods have been proposed.^{75–77} In this work, the simulation system was first subjected to 5000 steps of steep descent optimization. If there were crashes between peptide and lipid molecules, then the corresponding lipid molecules were removed from the system. Interestingly, crashes rarely occurred. The whole system was then pre-equilibrated for 1 ns with the peptide molecules constrained. The generated coordinates are used for simulation runs with the peptides fully relaxed at 323 K and 1 atm. To validate this simulation procedure, we compared the results with the self-assembling bilayer. Lipid molecules were randomly distributed in

Table 1. Experimental and Calculated Solvation Free Energies of UA Small Molecules in CG Hexadecane and CG Water

compound	solvation free energy in hexadecane (kJ/mol)			solvation free energy in water (kJ/mol) ^b		transfer free energy from hexadecane to water (kJ/mol)	
	CG	exptl ^a	experimental log P (hexadecane/gas)	CG	exptl	CG	exptl
ethane	-3.0	-2.8	0.5	7.3	7.4	10.3	10.2
propane	-6.2	-6.0	1.1	8.2	8.3	14.4	14.3
butane	-8.7	-9.3	1.6	9.2	9.0	17.9	18.3
methanol	-4.8	-5.6	1.0	-20.1	-20.2	-15.4	-14.6
ethanol	-8.1	-8.6	1.5	-20.5	-21.0	-12.4	-12.4
1-propanol	-10.6	-12.1	2.1	-20.1	-20.4	-9.5	-8.3
2-propanol	-10.4	-10.5	1.8	-22.0	-19.9	-11.6	-9.5
ethylamine	-9.3	-9.7	1.7	-19.4	-18.8	-10.1	-9.2
acetone	-10.3	-9.7	1.7	-14.1	-16.1	-3.8	-6.4
butanone	-12.3	-13.2	2.3	-15.2	-15.2	-2.9	-2.1
acetamide	-14.1	-14.0	2.4	-39.7	-40.6	-25.6	-26.6
benzene	-14.5	-16.0	2.8	0.0	-3.6	14.5	12.4
toluene	-21.8	-19.1	3.3	-3.7	-3.7	18.1	15.4
naphthalene	-26.7	-30.7	5.3	-7.2	-10.0	19.5	20.7
methylindole	-38.2	-	-	-31.0	-24.6	7.2	-
<i>p</i> -cresol	-28.7	-24.8	4.3	-29.0	-25.6	-0.3	-0.8
acetic acid	-10.1	-10.1	1.8	-27.4	-28.0	-27.4	-28.0
dimethyl sulfide	-13.2	-12.9	2.2	-8.5	-6.4	4.7	6.5
average error	1.0			1.1		1.1	

^a The solvation free energy in hexadecane is calculated from experimental partition coefficients between gas phase and hexadecane ($\log P$).^{78,79} ^b Ref 32.

the system box. The lipid bilayer then self-assembled around the peptides. The tilt angles of WALP23 in DOPC were calculated using both methods. We found that there was no significant difference in the tilt angle between the two methods (data not shown). This showed that small peptides could easily be fitted into the CG lipid model without the need for complicated setup procedures. Note that this probably applies to small simple peptides only. For large proteins, this simple scheme may not work.

RESULTS AND DISCUSSIONS

Parameterization of Protein–Lipid Tail Interactions. Each lipid molecule comprises two major parts: the nonpolar tail group and the polar headgroup. The nonpolar tail part of the lipid molecule is the core of the lipid bilayer. A large portion of transmembrane peptides is exposed to the nonpolar environment. This could be considered as solvating in the nonpolar solvent. One of the more promising parametrization strategies is to reproduce the experimental thermodynamic properties of organic molecules.^{19–21,32} To parametrize protein UA particles and CG lipid tail particles, we fitted the solvation free energies of UA small organic molecules in CG hexadecane, which is a better model than cyclohexane, for the membrane environment. The experimental partition coefficients ($\log P$) of small organic molecules between the gas phase and hexadecane are available.^{78,79} The solvation free energy can be calculated from the partition coefficient by:

$$\Delta G_{\text{sol}} = -2.303 RT \log P \quad (10)$$

where R is the universal gas constant and $T = 300$ K. Seventeen small molecules from eight classes of organic compounds

(alkane, alcohol, ketone, amide, amine, aromatics, carboxylic acid, and sulfide) covering different types of amino acid side chains were used to optimize the parameters. Ionizable amino acids are normally charged in an aqueous environment. But they may prefer to exist in the neutral form to avoid a large desolvation penalty when they are in a nonpolar environment. In the PACE force field, side chains of Asp, Glu, Lys, and Arg are assumed to be neutral inside the membrane but charged in the aqueous medium.³¹ Therefore, carboxylic acid was used for the calculation of the solvation free energies in hexadecane of Asp and Glu, while amine was used for Lys.

We first optimized parameters for aliphatic carbons, as these are the common part of all amino acid side chains. Parameters for $-\text{CH}_3$ were obtained from ethane, and $-\text{CH}_2$ could then be parametrized by propane and butane. Benzene was used to parametrize aromatic carbon. Then, other functional groups were parametrized using the corresponding small organic molecules. For simplicity, we adopted the Lorentz–Berthelot combination rule ($\delta_{ij} = (\delta_{ii} + \delta_{jj})/2$) to derive δ_{ij} . In hexadecane, ϵ_{ij} was optimized to reproduce experimental solvation free energy. The results and parameters are shown in Tables 1 and 2, respectively. Due to a lack of the experimental data, the S of Cys and Met side chains share the same parameters, so do the N of Lys and Arg side chains. Backbone N, the N of Asn, Gln, Trp, and His side chains share the same parameters. The absolute average error for the solvation free energies of UA small molecules in CG hexadecane is about 1.0 kJ/mol. The transfer free energies of these UA small molecules from CG hexadecane to CG water were obtained by calculating the difference between solvation free energy in water and in hexadecane:

$$\Delta G_{\text{hex} \rightarrow \text{water}} = \Delta G_{\text{water}} - \Delta G_{\text{hex}} \quad (11)$$

Table 2. Small Organic Molecules Used in the Calculation of Solvation Free Energy in Hexadecane for Protein–Lipid Tail Parameterization and the Resulting Optimized Parameters

protein UA group	compound used for parametrization	parameter	
		ϵ_{ij} (kJ/mol)	δ_{ij} (nm)
–CH ₃	ethane	1.7	0.43
–CH ₂	propane	1.3	0.43
	butane		
–CH	2-propanol	1.0	0.43
	benzene		
aromatic C ^a	toluene	0.9	0.4225
	naphthalene		
–N ^b	ethylamine	2.2	0.40
O=C–NH ₂ ^c	acetamide	2.1	0.40
–C=O ^d	acetone	1.0	0.415
–C=O ^d	butanone	2.0	0.375
–COO ^{–e}	acetic acid	1.0	0.415
–COO ^{–e}		1.7	0.375
–OH ^f	methanol	2.0	0.38
	ethanol		
–OH ^g	1-propanol	2.0	0.38
	<i>p</i> -cresol		
–S ^h	dimethyl sulfide	3.0	0.405

^a Aromatic C for Phe, Tyr, Trp and His side chains. ^b –N for Lys and Arg side chains. ^c –N for backbone amide, Trp, and His side chains. ^d –C=O for backbone amide, Asn, and Gln side chains. ^e –COO[–] for Asp and Glu side chains. ^f –OH for Ser and Thr side chains. ^g –OH for Tyr side chain. ^h –S– for Cys and Met side chains.

The absolute average error is about 1.1 kJ/mol. This shows that the PACE force field can reproduce the free energy of solvation and the free energy of partitioning between the oil and the aqueous phase well.

Parameterization of Protein–Lipid Head Interactions. It has been found that some of the amino acid side chains have favorable interactions with lipid head groups which influence the orientation of transmembrane peptides.^{80,81} Therefore, it is important to carefully parameterize the interactions between protein and lipid head groups. One of the successful parameterization strategies is to reproduce the partitioning free energies of amino acid side chains in a lipid bilayer using all-atom simulations. We have adopted this strategy for our parameter development.²¹ PMFs of amino acid side chain analogues in the DOPC bilayer were calculated and fitted with all-atom results.^{82,83} Figure 2 shows the PMFs of 18 amino acid side chain analogues (except glycine and proline) calculated using the OPLS-AA force field⁸² and the PACE force field. Table 3 gives the optimized parameters. Note that DOPC contains unsaturated CG carbon particles which are absent in CG hexadecane molecules. We tentatively assume that parameters for the interactions between UA particles and CG saturated carbon particles could be transferred to the CG unsaturated carbon particles in DOPC. The PMF results show that this assumption is valid in this case. Modification may be needed in the future.

Figure 1b shows the graphical representation of the DOPC system for PMF calculation. The far right of the PMF curves in Figure 2 indicates that the side chains stay in water and have very

little or even no interactions with the lipid bilayer. This corresponds to the solvation free energy of side chains in aqueous phase and is taken as the reference point. The far left of the curves indicates that the side chains stay at the center of the lipid bilayer. As the hydrophobic thickness of the DOPC bilayer is large (2.96 nm determined experimentally),⁸⁴ the side chains can be considered to be solvated in the nonpolar environment. As the last point of the PMF curve is taken as the reference point, the first point of the PMF curve could be regarded as the transfer free energy of the side chain from water to the hydrophobic environment. The accuracy of the calculated transfer free energies indicates the transferability of our parameters. The results show that the transferability of our model is reasonably good and can be transferred to different types of lipid bilayer. The middle region of the PMF curves shows the free energy of interactions between the side chain and lipid head groups relative to the reference state. To obtain correct descriptions of protein–lipid head interactions, the free energy profiles of this region must be accurately fitted.

We used the similar strategy as the parameterization of protein–lipid tail interactions. Combination rule was used to obtain δ_{ij} and ϵ_{ij} was optimized by fitting the all-atom PMFs. Parameters for –CH₃ were obtained from PMF of Ala. Then, PMFs of Val and Ile were used for the parameterization of –CH₂. Parameters of –CH and aromatic C could be optimized in a similar way using Leu and Phe, respectively. Other functional groups were parameterized by fitting PMFs of all amino acid side chain analogues. Lipid head composes of three CG particles: glycerol, phosphate, and choline. As they are at different positions of the lipid bilayer, we found that change of the parameters would have a different effect to the PMF curves. Phosphate is at the center of the PMF curve, the PMF at ~0 nm is the most sensitive to the change of the parameters. Glycerol is nearer to the lipid tail than phosphate. It mainly affects the PMF at about –0.5 nm. Choline is closest to water, and the effect of the parameters is mainly reflected at ~0.5 nm of the PMF. Interactions between UA particles and these three CG particles were parameterized by carefully fitting the whole PMF curves.

Alkane compounds have a high transfer free energy from hexadecane to water (~10–18 kJ/mol in Table 1). It is expected that hydrophobic residues, Ala, Val, Ile, and Leu prefer to stay in the lipid core. Our results are in very good agreement with the all-atom results. When the size of the hydrophobic side chains increases, the interaction with nonpolar lipid tails increases, and the free energy difference becomes larger. The free energy difference of hydrophobic residues between water and the center of the bilayer is about 8.4 to 22.1 kJ/mol. CG results for Ala, Val, and Leu are in excellent agreement with the all-atom results. The deviation is about 2 kJ/mol for Ile. Although both the side chains of Leu and Ile consist of four carbon atoms with the same interaction parameters with lipid tails, the relative free energy at the center of bilayer for Ile is 7 kJ/mol more favorable than that for Leu. This is because the branched Leu side chain packs into the lipid bilayer less efficiently than the linear Ile side chain. This effect can be observed in our model where every heavy atom of amino acids is represented explicitly. This phenomenon is absent when these side chains are represented by the same CG particles, indicating the need for finer protein models to describe the interactions between proteins and lipid heads. The free energy barrier at the headgroup region in our model is present in these cases. The deviation from the all-atom results is about 1 kJ/mol for Val, Ile, and Leu. The barrier is underestimated by about

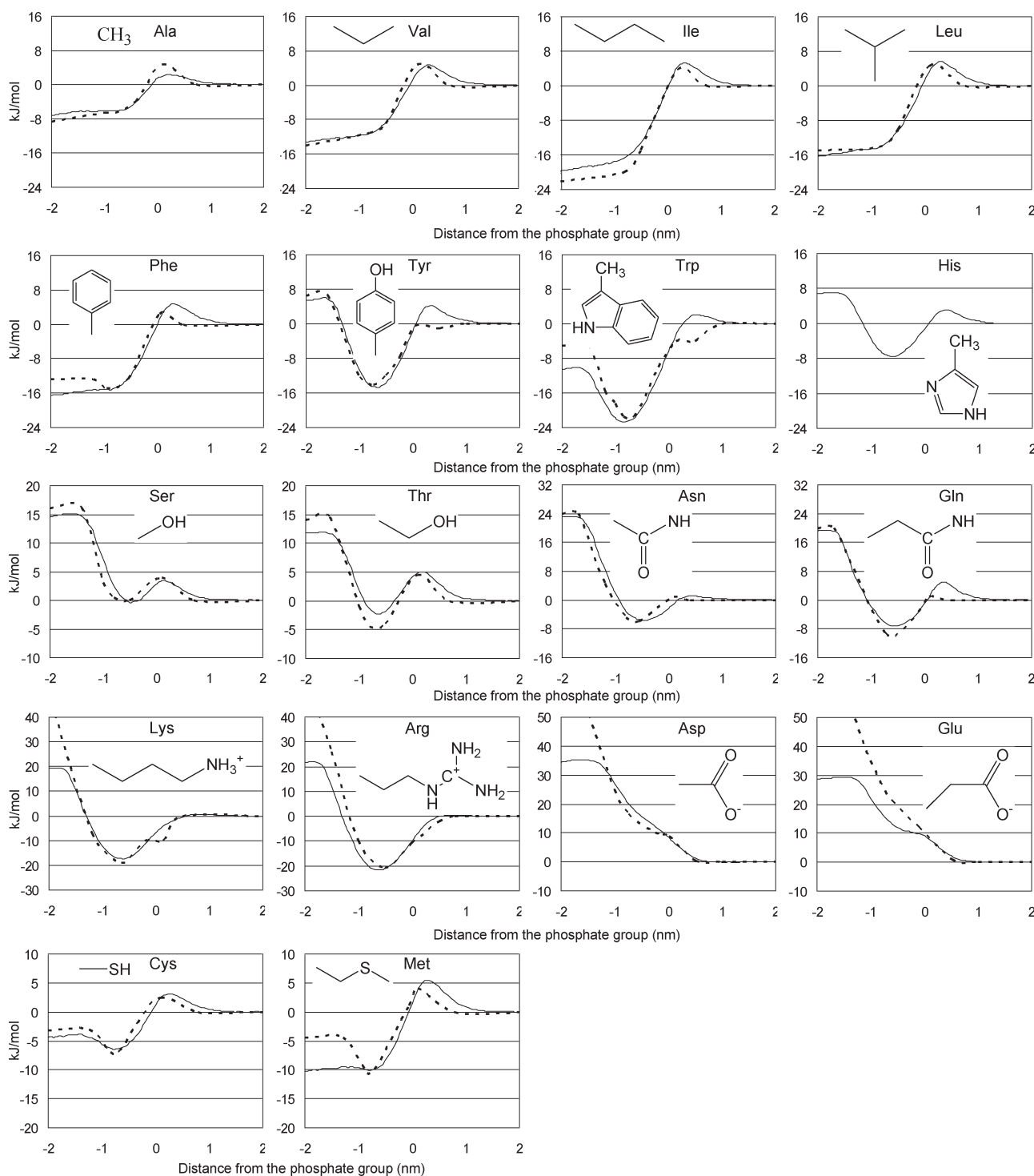


Figure 2. PMFs of 18 amino acid side chain analogues. CG results and all-atom results⁸² are shown in solid lines and dotted lines, respectively.

2.5 kJ/mol for Ala. Note that in the case of Ala, a single UA particle of CH₃ instead of CH₄ is used for parametrization. This may account for the small deviation of the PMF curves of Ala.

For aromatic residues, Tyr, Trp, and His have strong interactions with lipid head groups, as they are able to form H-bonds with lipid head groups. The minima for Tyr and Trp are well reproduced. The relative free energies at the center of the bilayer are reasonably reproduced for Tyr. For Phe and Trp, our model shows 4–5 kJ/mol more stabilization than the all-atom model.

All-atom simulations for the solvation free energy calculations of 15 neutral amino acid side chains using the AMBER(FF94), CHARMM22, and OPLS-AA force fields show an average error of 4.6–6.8 kJ/mol.⁸⁵ All-atom calculations of transfer free energies from hexadecane to water for all amino acid side chains using the OPLS-AA force field have an average error of about 4 kJ/mol.⁸⁶ Considering the errors associated with the solvation parameters of all-atom force fields, no further optimization was done to minimize the transfer free energy difference between our

Table 3. Optimized Parameters for Protein–Lipid Head Interactions

protein UA group	CG lipid headgroup					
	glycerol (N _a)		phosphate (Q _a)		choline (Q _b)	
	ϵ_{ij} (kJ/mol)	δ_{ij} (nm)	ϵ_{ij} (kJ/mol)	δ_{ij} (nm)	ϵ_{ij} (kJ/mol)	δ_{ij} (nm)
–CH ₃	1	0.43	0.4	0.43	0.4	0.43
–CH ₂	1	0.43	0.4	0.43	0.4	0.43
–CH	0.4	0.43	0.4	0.43	0.4	0.43
aromatic C ^a	0.7	0.4225	0.7	0.4225	0.7	0.4225
–N ^b	10	0.4	17	0.4	15 ^k	0.4
O=C–NH ₂ ^c	2.1	0.4	2.1	0.4	2.1	0.4
O=C–NH ₂ ^c	5	0.375	5	0.375	0	0.375
–C=O ^c	1	0.415	0.4	0.415	0.4	0.415
–C=O ^c	2	0.375	20 ^k	0.375	2	0.375
–NH ^d	2.1	0.4	2.1	0.4	2.1	0.4
–NH ^d	6	0.375	6	0.375	0	0.375
–C=N ^e	2.1	0.4	2.1	0.4	2.1	0.4
–N ^f	6	0.4	6	0.4	15 ^k	0.4
–C ^{fg}	1	0.415	6	0.415	15 ^k	0.4
–COO ^g	1	0.415	0.4	0.415	0.4	0.415
–COO ^g	1.7	0.375	20 ^k	0.375	7	0.375
–OH ^h	6	0.38	2	0.38	2	0.38
–OH ⁱ	8	0.38	2	0.38	2	0.38
–S ^j	4	0.405	3	0.405	3	0.405

^a Aromatic C for Phe, Tyr, Trp and His side chains. ^b –N for Lys and Arg side chains. ^c –C=O and O=C–NH₂ for backbone amide, Asn, and Gln side chains. ^d –NH for Trp and His side chains. ^e –C=N for His side chain. ^f –N and –C⁺ for Arg side chain. ^g –COO[–] for Asp and Glu side chains. ^h –OH for Ser and Thr side chain. ⁱ –OH for Tyr side chain. ^j –S– for Cys and Met side chains. ^k The following potential is used: $E_{\text{nonbonded}} = \sum_{i \neq j} (4\epsilon_{ij}\delta_{ij}^{12})/r^{12}$.

model and the all-atom model. Our model shows free energy barriers of about 2–4.5 kJ/mol but only about 2.5 kJ/mol (Phe) or even no barrier (Tyr and Trp) in the all-atom results. The existence of the barrier may be due to the lipid model, as it is also found in the MARTINI model. There is room for improving our force field. As there is no all-atom result for His, a comparison cannot be made.

For polar residues, Ser, Thr, Asn, and Gln each contain a hydroxyl group or amide group that is able to form H-bonds with lipid head groups. Our model can reproduce the minima of these residues at the lipid head region accurate to ± 3 kJ/mol. The heights of the barrier for Ser and Thr are in very good agreement with the all-atom results. Asn shows no barrier in both our model and the all-atom model. But there is a barrier for Gln in our calculation that is absent from the all-atom simulation.

For all-atom PMFs, the relative free energies of ionizable residues at the center of the bilayer is very large (>50 kJ/mol). This is because charged molecules have large desolvation penalties when they are moved from the aqueous medium to the nonpolar environment. Our results show smaller relative free energies at the center of the bilayer as these residues are considered to be neutral in the membrane, but they were assumed to be charged in the all-atom calculations. Tieleman et al. calculated pK_a values of ionizable groups in DOPC using all-atom model. Asp and Glu were found to be neutral in the lipid core, whereas Lys and Arg prefer charged state.⁸² It indicates that our assumption may be valid for Asp and Glu only. Introduction of charged Lys and Arg will be considered in the future. The ionizable residues are more likely to enter the lipid core in our model than in the all-atom models. The middle region of the PMF curves,

which was our main focus and indicates the interactions between protein particles and the lipid head, was fitted to the all-atom results. For Lys and Arg, the free energy minima indicating strong polar–polar interactions with lipid head groups are well reproduced. In the case of negatively charged side chains, the PMF of Asp was fitted to the all-atom result, and the optimized parameters were transferred to Glu. Both residues show unfavorable interactions with the lipid bilayer. However our model is somewhat less unfavorable for Glu in the middle region than the all-atom calculation. This may indicate that Asp and Glu should not share the same parameters and that further improvement should be made in the future.

Both the PMFs of Cys and Met match the all-atom results except that Met shows more favorable interactions with the lipid tail in our model than in the all-atom model. Our model shows that Met is 10 kJ/mol more stable in the lipid tail, whereas it is only 4.4 kJ/mol more stable in the lipid tail in the all-atom calculation. The experimental transfer free energy of Met from cyclohexane to water is 9.8 kJ/mol.⁸⁷ These results show that our model is comparable to the all-atom simulations in describing the partitioning of amino acid side chains in the lipid bilayer.

Modification of Backbone Hydrogen-Bonding Potential in Membrane Environment. Before we present the detailed results obtained with the optimized parameters for lipid–protein interactions, the hydrogen-bonding (HB) potential of PACE needs to be further investigated. The original HB parameters of PACE were obtained by reproducing experimental α -helical and β -sheet contents of model peptides in aqueous simulations.³¹ As the simulations were carried out with the MARTINI water model that bears no dipole moment, the screening effect of water was

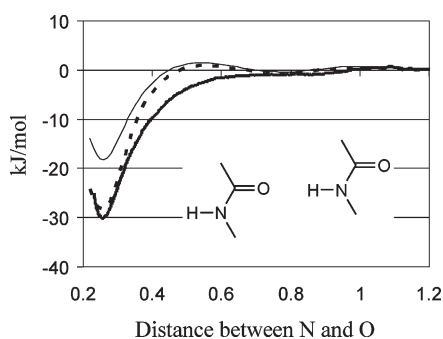


Figure 3. PMFs of two amides in pure octane. CG calculations using original and modified HB potentials and all-atom calculation are shown in thin, dotted, and thick lines, respectively.

implicitly taken into the original HB parameters. Hence, the electrostatic part of HB may be underestimated in membrane environment. We tried to estimate the underestimation of the electrostatic part by calculating the dimerization free energy of *N*-methylacetamide in octane with both the original HB parameters of PACE and OPLS-AA.⁸⁸ As shown in Figure 3, the HB parameters of PACE lead to the dimerization energy of 18.2 kJ/mol in octane, whereas OPLS-AA gives dimerization energy of 29.5 kJ/mol. To account for the difference between the PACE and OPLS-AA results, we generated a modified PACE in which ϵ_{attr} (eq 8) was increased so that the dimerization energy of *N*-methylacetamide in octane was reproduced.

To examine the effect of the modification, we performed the simulations of WALP-based peptides and glycoporphin A helix dimer in membrane with both the original and modified PACEs. The comparison reveals that all the WALP-based peptides have stable helices and similar tilt angles for both HB potentials of PACE (Table S1, Supporting Information). For the glycoporphin A helix, the helical structure in the region of GVIG becomes unstable in the simulation with the original PACE, owing to the helix-breaking propensity of the two glycine in the middle of helix. The helical structure in the same region is however stable in the simulation with the modified PACE. The comparison suggests that the original PACE can be used to simulate the membrane helical peptides without glycine in the middle of helix. For the peptide like the glycoporphin A helix, the modified PACE is needed. In the following sections, only the results in the simulations with the modified PACE will be discussed in detail.

It should be noted that the major limitation of our way of modifying the HB parameters is that all HB will be strengthened, although the dielectric environment across membrane is quite different⁸⁹ and thus HBs located at different places of membrane should have distinct strength. Moreover, when HBs dynamically change their positions during simulations, their strength should also change accordingly. Therefore, a more desirable model should be able to adjust HB strength on the fly, which is under our investigation. Alternatively, a recent polarizable CG water model in MARITNI may also be a good choice.⁹⁰

Tilting of Helical Peptides. Different transmembrane peptides show significant variations in the tilt angle in lipid bilayers. For example, the influenza A M2 channel was reported to have a tilt angle of 38°, whereas the channel-lining M2 segments from the d-subunit of the nicotinic acetylcholine receptor shows a tilt angle of 14°. ^{91,92} Artificial transmembrane peptides were designed to systematically study the factors affecting the orientation of transmembrane peptides. It is generally accepted that

Table 4. Amino Acid Sequences of the Peptides Used in the Literature and in Simulation

peptide	sequence
WALP23	acetyl-GWW(LA) ₈ LWWA-amide
WALP19	acetyl-GWW(LA) ₆ LWWA-ethanolamine
WALP19-P10	acetyl-GWW(LA) ₃ P(AL) ₃ WWA-ethanolamine
GWALP23	acetyl-GGALW(LA) ₆ LWLAGA-amide
KWALP23	acetyl-GKALW(LA) ₆ LWLAKA-amide

hydrophobic mismatch affects the orientations and therefore functions of membrane proteins. When the hydrophobic length of the peptide is larger than the hydrophobic width of the lipid (positive mismatch), the peptide tilts to allow it to have better interactions with the lipids. When the hydrophobic length of the peptide is smaller than the hydrophobic width of the lipid (negative mismatch), the tilting of the peptide is smaller.⁹³

In this work, we calculated the tilt angles of WALP and GWALP peptides and their mutants. Table 4 shows the amino acid sequences of the peptides that were studied. WALP peptides were chosen because they have been widely studied in experiments and computer simulations.^{34–43} It is interesting to compare our results with theirs. We found that the dynamic model should be used instead of the static model to determine the tilt angle. We also studied GWALP and KWALP peptides as they have recently been examined using the dynamic model.⁴⁴ Moreover, we report the first simulation result of WALP19-P10 which shows a different tilting preference from WALP19.⁴⁵

The peptides were vertically inserted into the pre-equilibrated lipid bilayers with initial fully α -helical structures. Each simulation was performed for 300 ns, and the last 250 ns were used for analysis. The average tilt angles of the simulated peptides and the calculated hydrophobic widths of lipids and lengths of peptides are shown in Table 5. The tilt angle of a peptide was calculated as the angle between the helical axis of the hydrophobic segment of the peptide and the normal of the lipid bilayer. As the bilayer normal was aligned along the *z*-axis, the *z*-axis was taken so there was no need to calculate the bilayer normal. The hydrophobic width of the lipid bilayer is defined as the average distance between the first hydrophobic beads in the two leaflets. The peptide hydrophobic length is the distance of all the hydrophobic residues between the anchoring groups. For ideal helical peptides, the hydrophobic length for each residue is 0.15 nm. The deviations of the calculated hydrophobic width of lipids and length of peptides from experimental data were smaller than 0.05 nm. The calculated tilt angles of WALP19 and WALP23 were in the range of 7.5° to 17.5° which shows the hydrophobic mismatch effect. The tilting of WALP19 was determined using the static model in experiment. We found a larger tilting than experimental results. This suggests that the dynamic model should be used. GWALP and KWALP also show a hydrophobic mismatch effect, and the calculated tilt angles are in good agreement with experimental results which used the dynamic model. Interestingly, WALP19-P10 did not show a hydrophobic mismatch effect and has the same tilting in both DOPC and DLPC. Our simulation supports the experimental finding that the Pro10 affects the tilting.

One of the most extensively studied artificial transmembrane peptides is WALP.^{34–43} WALP is a poly-(Leu-Ala) peptide with two Trp residues at both ends. It has been used to develop and validate CG force fields.^{21,101} The exact tilting of helical peptides

Table 5. Average Tilt Angles of WALP, GWALP Peptides and Their Mutants and Calculated Hydrophobic Widths of Lipids and Lengths of Peptides^a

peptide	tilt angle (°)				calculated lipid hydrophobic width ^g (nm)		calculated peptide hydrophobic length ^h (nm)	
	DOPC		DLPC		DOPC	DLPC	DOPC	DLPC
	sim.	exptl.	sim.	exptl.				
WALP23	7.5 (5.0)	11 ^c	17.5 (7.6)	15 ^d	3.01 (0.05)	1.93 (0.04)	2.50 (0.03)	2.49 (0.03)
WALP19	8.0 (5.7)	4.0 ^e	13.5 (7.2)	4.0 ^e	2.99 (0.05)	1.93 (0.04)	1.94 (0.02)	1.94 (0.02)
WALP19-P10 ^b	15.0 (8.0)	11.6 ^e	15.5 (8.8)	11.9 ^e	2.99 (0.05)	1.92 (0.05)	1.97 (0.04)	1.97 (0.04)
tilt difference	7.0	7.6	2.0	7.9				
GWALP23	6.5 (5.3)	6.0 ^f	15.5 (7.5)	18.6 ^f	2.99 (0.05)	1.93 (0.04)	1.94 (0.02)	1.94 (0.02)
KWALP23	7.5 (5.7)	7.3 ^f	18.0 (8.0)	18.0 ^f	3.00 (0.05)	1.93 (0.04)	1.94 (0.02)	1.94 (0.02)
tilt difference	1.0	1.3	2.5	-0.6	exptl value: 2.96 ⁱ	exptl value: 1.95 ^j		

^aAll the experimental tilt angles were determined using the dynamic model except for WALP19 and WALP19-P10. Standard deviations are shown in parentheses. ^bAverage tilt is calculated for the C-terminal segment of WALP19-P10. ^cRef 43. ^dRef 98. ^eRef 45. ^fRef 44. ^gThe lipid hydrophobic width is the average distance between the first hydrophobic beads in the two leaflets. ^hThe peptide hydrophobic length is the distance between the anchoring groups for all the hydrophobic residues. For ideal helical peptides, the hydrophobic length for each residue is 0.15 nm. ⁱRef 84. ^jRef 113.

is still debated, as experimental and simulation results differ greatly.⁹⁴ WALP23 was found to have a tilt angle of about 12° in dimyristoylphosphatidylcholine (DMPC) by ATR-FTIR spectroscopy.⁹⁵ Recent studies by fluorescence spectroscopy found large tilt angles for WALP23 in DOPC (23.6°).⁹⁶ Koeppe et al. developed another method that uses solid-state ²H NMR based on geometric analysis of labeled alanines (GALA) to study the orientation of a transmembrane peptide. This technique allows a higher resolution of tilt angles (<1°), but small tilt angles were determined for WALP23 in DOPC (4.8°) and DLPC (8.1°).^{38,80}

The small angles found by the GALA method have been explained by computer simulation studies.⁹⁷ In the conventional GALA method, the tilt angle of a peptide is calculated by fitting two parameters, τ (tilt angle) and ρ (rotation angle) to the experimentally measured quadrupolar splitting of labeled alanines. Fluctuation of the peptide leads to an averaging effect so the tilt angle would be underestimated. To take the fluctuation of the peptide into account, the global order parameter S could be used as a fitting parameter instead of taken as a constant. With this parameter, the effects of additional internal vibrations, rotations, and wobbling of the peptide are included (model 3 in ref 98). Tilt angles of WALP23 in DMPC and DLPC have been found to be 7° and 15°, respectively.⁹⁸ This dynamic model has been used in recent studies to determine the tilt angles of the GWALP peptide and its mutants.^{44,81} Strandberg et al. proposed to explicitly consider the fluctuations of τ and ρ by introducing the Gaussian distribution to them. A total of four parameters including two additional parameters, σ_τ and σ_ρ , were used to fit the experimental data (model 6 in ref 98). Larger tilt angles were obtained for WALP23 in DMPC (14°) and DLPC (29°).⁹⁸

All-atom MD simulations show larger tilt angles compared with the results from solid-state ²H NMR. In a study using the CHARMM22 force field with the GBSW implicit membrane, tilt angles of 32.7° and 15.5° were found for WALP23 and WALP19, respectively, in a 0.23 nm thick membrane hydrophobic core that corresponds to DMPC.⁴¹ A similar result was obtained when the ffgmx force field was used with the explicit membrane (33.5° for WALP23 in DMPC).⁹ Slightly smaller tilt angles were determined for WALP23 (28.1°) and WALP19 (12.1°) in DMPC with the CHARMM22 force field and the explicit

membrane.^{99,100} A tilt angle of 14° was found for WALP23 in DPPC using the MARTINI CG force field and Bond and Sansom's protein model. But these models showed larger tilting for shorter peptides (22° for WALP19).¹⁰¹ Another study using the MARTINI protein and lipid model showed a tilt angle of 11.4° in DOPC and 23.7° in DLPC for WALP23.¹⁰²

In our work, WALP23 shows moderate tilting in DOPC (7.5°) and DLPC (17.5°). The results are in good agreement with that determined by the GALA method using the dynamic model. This supports the argument that the dynamic motion of the peptide should be considered. But the values are smaller than those obtained using the Gaussian distribution. The question of whether the Gaussian distribution is suitable for describing the motion of peptides has been raised.¹⁰² The fluctuations of tilt and rotation angles are assumed to follow a Gaussian distribution. But both all-atom simulations and long CG simulations showed non-Gaussian distributions. In particular, the rotation angle was found to follow a distribution that is far from the Gaussian one.^{81,102,97} Our results do not show a strict Gaussian distribution for the tilt angle either (Figure 4). When a uniform distribution was applied instead of a Gaussian distribution, a larger tilt angle was found for WALP23 in DMPC (21° rather than the 14° in ref 98).¹⁰³ The most suitable functional form to describe the motion of peptides has not yet been determined. Our results suggest that introducing S as a free parameter is enough to account for the motion of peptides.

WALP19 in DOPC and DLPC was found to have tilting angles of 8.0° and 13.5°, respectively (Table 5). The GALA method without considering the dynamic motion of peptides showed smaller tilt angles in DOPC and DLPC (4°). WALP19 and WALP23 in DOPC show small tilt angles as expected under negative mismatch. WALP19 in DLPC is under the hydrophobic match condition. As results determined by the dynamic model are unavailable, we suspect that WALP19 may have a tilt angle of about 13.5° in DLPC. Similar tilt angles were also obtained for peptides under the hydrophobic match condition by all-atom (KALP23 in DMPC) and CG (WALP23 in DPPC) simulations.^{93,102} The hydrophobic mismatch effect is mainly due to favorable helix-lipid interactions.¹⁰⁰ The anchoring groups of WALP at the terminals, Trp, were mainly situated at the lipid head region. This is because Trp side chains interact favorably

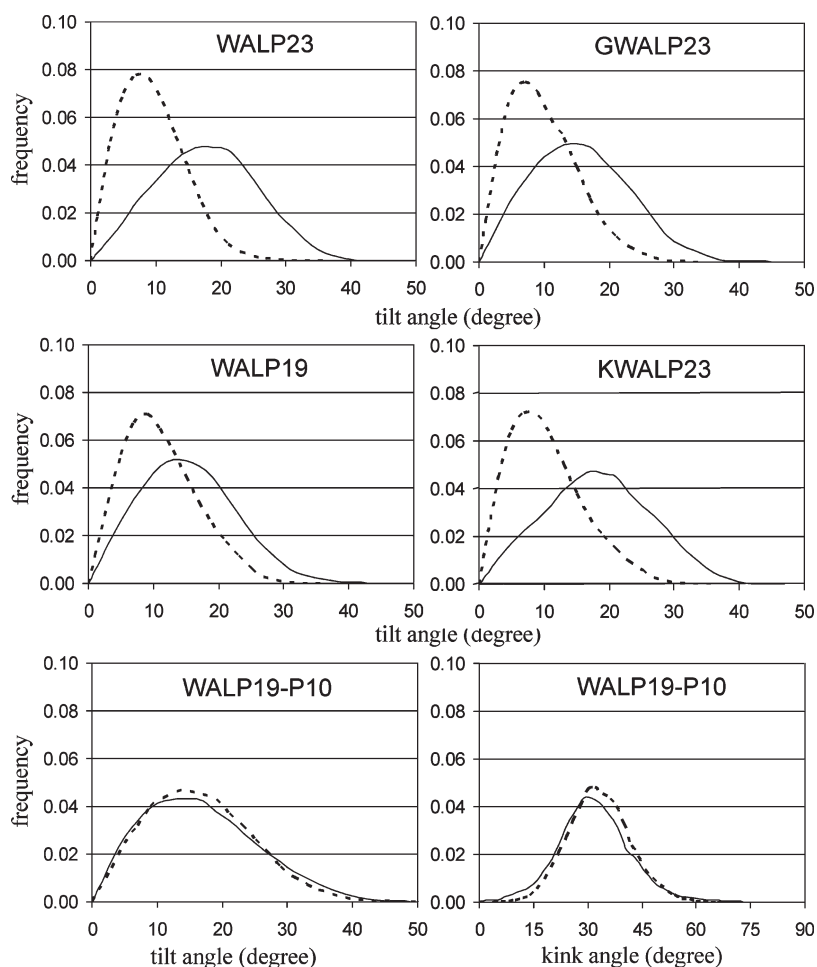


Figure 4. Distribution of tilt and kink angles of WALP and GWALP peptides and their mutants in the DOPC (dotted line) and DLPC (solid line) bilayers.

(22 kJ/mol) with lipid head groups. (Figure 2) Therefore, WALP23 has a larger tilt angle in DLPC than WALP19 in order to position the anchoring groups at the lipid head region.

The effect of hydrophobic mismatch is also observed for another family of peptide, GWALP. Unlike WALP peptides, GWALP peptides contain only one Trp residue at both terminals and positioned at the inner region. Tilt angles of GWALP23 and its mutant, KWALP23, in DOPC and DLPC were determined using a dynamic model similar to model 3 in ref 98. The tilt angle of GWALP23 in DLPC was analyzed by fitting ^2H quadrupolar splittings and $^1\text{H}-^{15}\text{N}$ dipolar coupling data. Tilt angles of 6.0° and 18.6° were found for GWALP23 in DOPC and DLPC, respectively.⁴⁴ Our results match the experimental findings very well. There is no significant difference in the tilt angle between KWALP23 with an Lys residue at the two ends and GWALP23.⁴⁴ Lys is also an anchoring group that prefers to position at the lipid interface and has been shown to affect the tilting of the transmembrane peptides.⁸⁰ In the case of KWALP23 containing both Lys and Trp residues, inner Trp seems to determine the tilting but outer Lys does not. Our results are in excellent agreement with the experimental data for KWALP23 and also support this finding from the experiment.

Apart from the studies of artificial transmembrane peptides with Leu and/or Ala in the hydrophobic core, mutation at the center of the peptides has been carried out to study the effect of a

particular amino acid on the orientation of the peptide.^{45,81} WALP19-P10 is a mutant of WALP19 with Leu10 replaced by Pro. Pro has no backbone amide hydrogen and is a helix-breaking residue. It was confirmed by circular dichroism spectroscopy that the helicity of WALP19-P10 is smaller than that of WALP19.⁴⁵ Our simulation result shows that the helix is broken at Pro10 and that two helical segments were observed instead of one. (Figure 5) A tilting angle of $\sim 12^\circ$ was found for the C-terminal segment in both DOPC and DLPC by solid-state ^2H NMR spectroscopy.⁴⁵ A tilt angle of $\sim 15^\circ$ was obtained in our calculations (Table 5). The tilting of WALP19-P10 in DOPC is apparently larger than that of WALP19. This indicates that the kink at Pro10 affects the tilt angle of the C-terminal segment. A kink angle of $\sim 19^\circ$ was found experimentally for WALP19-P10.⁴⁵ Our simulation found average kink angles of 32° and 31° in DOPC and DLPC, respectively (Figure 4). Kink angles ranging from 9° to 41° were found for 48 Pro-containing helices in the crystal structures of soluble proteins.¹⁰⁴ Kink angles of $0^\circ-70^\circ$ were observed for 50 transmembrane helices containing proline in the crystal structures of membrane proteins.¹⁰⁵ Upon using the static model for the analysis of WALP19-P10 in the experiment, a kink angle of $\sim 19^\circ$ was suggested to be the lower limit.⁴⁵ Our result suggests that the kink angle may be as large as $\sim 30^\circ$ and that the kink also affects the tilting preference.

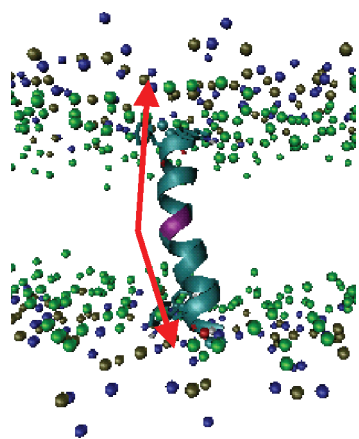


Figure 5. Snapshot of the WALP19-P10 peptide simulation in the DLPC bilayer. The residue Pro10 is shown in purple. The helical axes of N and C terminal segments are shown by the upper and lower arrows, respectively. The choline, phosphate, and glycerol particles of the DLPC molecules are shown in blue, brown, and green respectively.

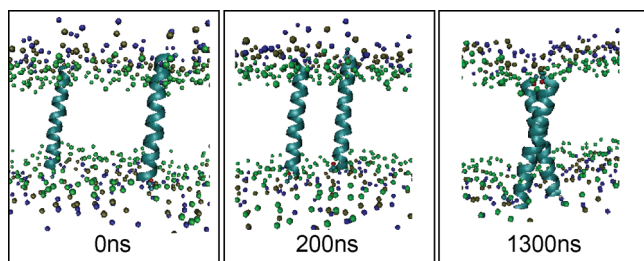


Figure 6. Snapshots of GpA helix associations in the DPPC bilayer at 0 ns, 200, and 1300 ns. The initial configuration of two GpA helices has a separation of ~ 3.5 nm. The choline, phosphate, and glycerol particles of the DPPC molecules are shown in blue, brown, and green, respectively.

Association of Glycophorin A Helix. Helix bundles are commonly found in membrane proteins. Understanding the folding and assembling of them enables us to predict their structures and to design new membrane proteins.^{106,107} The two-stage folding model has been proposed for the assembling of transmembrane helices: Each transmembrane helix is inserted into the membrane independently followed by the association of helices.¹⁰⁸ In this work, we studied the association of the GpA helix. The GpA dimer contains a seven-residue motif ($L^{75}I^{76}xxG^{79}V^{80}xxG^{83}V^{84}xxT^{87}$) that has been found to be important in the packing and dimerization of GpA helices.¹⁰⁹ Previous CG simulations have attempted to study such processes.^{67,68} Unfortunately, although a stable dimer was observed, significant deviations of the simulated structure from the experimental one were evident. It would be interesting to use the hybrid-resolution model to study such a system to validate our force field.

Two GpA helices (acetyl-EITLIIFGVMAVGITILLISYGIR-methylamide) were inserted into DPPC in a parallel fashion with a interhelix separation of ~ 3.5 nm. (Figure 6) After a 1 ns pre-equilibrium simulation with a position constraint on the peptides, the two helices were allowed to move freely in the system box for 2 μ s. Three individual trajectories with different initial velocities were obtained, with a total of 6 μ s of simulations. Interhelix separation, defined as the distance between the center

Table 6. RMSD from NMR Structure in DPC Micelle and the Crossing Angle of GpA Dimer in Three Runs

GpA simulation	$C\alpha$ RMSD (nm) ^a		crossing angle Ω ($^\circ$)	
	all residues	seven-residue dimerization motif ^b	our work	expt
run 1	0.228 (0.051)	0.203 (0.045)	-37.0 (7.8)	$-40^\circ, -35^\circ$ ^d
run 2	0.294 (0.032)	0.280 (0.029)	-43.5 (6.1)	
run 3	0.296 (0.092)	0.260 (0.052)	-32.5 (8.7)	
concatenated	0.272 (0.071)	0.249 (0.059)	-35.8 (9.8)	

^a Average RMSDs were calculated relative to the DPC micelle NMR structure.⁴⁸ ^b Only seven residues in the dimerization motif (LIxxGVxx-GVxxT) were considered in the calculation. ^c Derived from solution NMR structure in DPC micelle.⁴⁸ ^d Derived from solid-state NMR structure in the DMPC bilayer.⁵² Standard deviations are shown in parentheses.

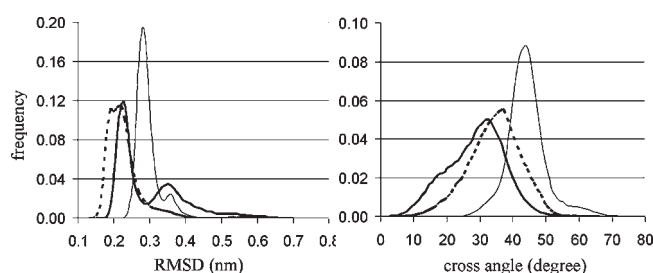


Figure 7. Distributions of RMSD relative to the NMR structure in DPC micelle and the crossing angle of GpA dimer in three runs. RMSD calculations were performed for all residues. Runs 1–3 are represented by dotted, thin, and thick lines, respectively.

of mass of the two GpA helices, as a function of time for three runs is shown in Figure S2, Supporting Information. The two helices met one another at ~ 200 ns, after which no disassociation was observed throughout the simulations. This indicates that the helix dimer is stable and is consistent with experiments and other CG simulations.^{53,67,69,110,111} The ensemble after 300 ns of simulation was used for the analysis of the structural features of the GpA dimer in each run. Thermodynamic properties were not evaluated, as the trajectories had not fully converged.

Table 6 shows the $C\alpha$ root-mean-square-deviation (RMSD) of the simulated dimers from the solution NMR structure in DPC micelle. The distributions of $C\alpha$ RMSDs considering all residues for all three runs are shown in Figure 7. The most probable RMSDs are ~ 0.22 nm for two of the runs and ~ 0.28 nm for another run. On average, $\sim 14\%$ of all the simulated structures have $\text{RMSD} \leq 0.2$ nm, and $\sim 63\%$ of structures have RMSD between 0.2 nm and 0.3 nm. When the seven-residue dimerization motif was considered only in the calculation, the $C\alpha$ RMSDs were reduced by 0.014–0.036 nm, with the RMSD of the concatenated trajectory shrinking to just 0.249 nm. These values are smaller than that obtained using the MARTINI CG protein force field (0.36 nm).⁴⁷ The standard deviation for concatenated trajectory (0.071 nm for all residues and 0.059 nm for the seven-residue dimerization motif) was also lower than that obtained using the MARTINI CG protein force field (0.13 nm). Our model may give a better structural representation of proteins in membrane (Figure 8).

Another well-defined structural feature of the GpA dimer is the crossing angle between the two helices. The GpA dimer has

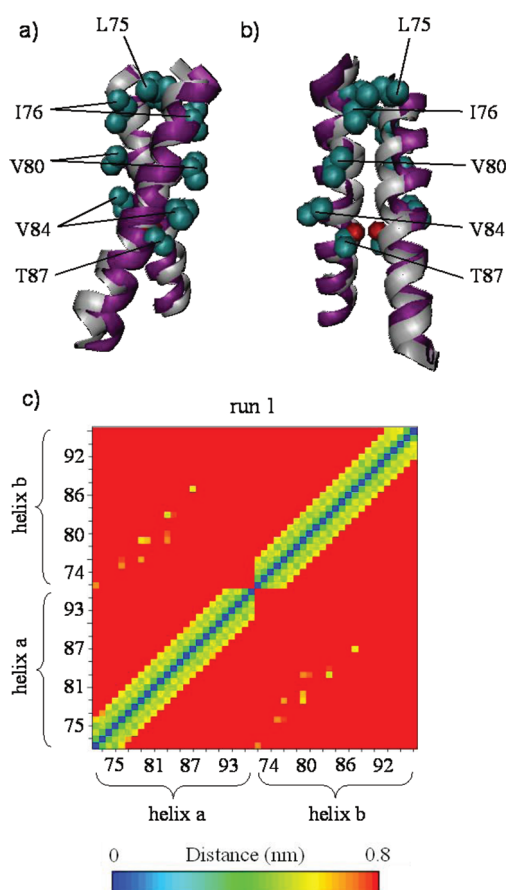


Figure 8. Representative structure of the simulated GpA dimer in purple in: (a) front and (b) side views. Side chains of key residues, L75, I76, V80, V84, and T87, are represented by green (hydrophobic particles) and red (hydrophilic particles) spheres. NMR structure in micelle is shown in gray for reference. (c) Contact map of the GpA dimer for the first run. The matrices correspond to the ensemble of the simulations with the cutoff distance of 0.8 nm. Seven common contacts were found in the three runs: L75–I76, I76–G79, G79–G79, G79–V80, G79–G83, G83–G83, and T87–T87.

right-handed packing with a negative crossing angle. The experimental crossing angles were -40° and -35° , as determined by solution NMR for DPC micelle and solid-state NMR for DMPC bilayer, respectively.^{48,52} Different media may perturb the structure of the GpA dimer leading to slightly different values obtained in the micelle and lipid bilayer. The average crossing angles for the three runs were -37.0° , -43.5° , and -32.5° (Table 6). They are all consistent with the experimental findings. The result (-35.8°) of the concatenated trajectories is in excellent agreement with the experimental value determined for lipid bilayer (-35°). This is supported by all-atom simulations which find that the crossing angle in the lipid bilayer is smaller than that in micelle by 3° – 7° .⁶⁰ The distributions of the crossing angle are shown in Figure 7. The distributions are larger than those obtained in all-atom simulations. The standard deviation of the crossing angle is 9.8° for the concatenated result and 2° – 5° for all-atom results using the GROMOS force field.⁶⁰ This may be due to the short simulation length used by the all-atom force field (50 ns) or the less rugged free energy landscape of the CG force field. Large distributions of the crossing angle were also found using the MARTINI protein force field. The average crossing angle

determined by previous CG simulation was -20° to -25° , smaller than the experimental values by 10° – 15° .^{67,69} Moreover, a positive crossing angle not observed in NMR structures was sampled. This indicates the limitation of the current CG protein force field and the higher accuracy and applicability of the PACE force field in studying helix assembly.

To further evaluate the structural features of the simulated GpA dimer, the contact maps for three runs were calculated and are shown in Figure 8c and Figure S3, Supporting Information. Seven common contacts were found in the three runs: L75–I76, I76–G79, G79–G79, G79–V80, G79–G83, G83–G83, and T87–T87. Residues involved in these contacts are in fact all the key residues of the $L^{75}I^{76}xxG^{79}V^{80}xxG^{83}V^{84}xxT^{87}$ motif. The contact analysis showed that binding of helices is mainly due to interactions among these residues. In a statistical study of the most frequently occurring motif that mediates helix–helix interactions, 13 606 transmembrane domains were analyzed.¹¹² GxxxG was found to be the most significant motif. Among seven common contacts found in our simulations, five of them involve Gly. This is because Gly has no side chain. This less bulky residue allows closer packing of helices. Even though Val84 was not one of the common contacts in the three runs, it is nevertheless one of the contacts in the runs 1 and 2. Previous simulations only showed three contacts (L75–I76, G83–G83, and T87–T87) using the same cutoff distance.⁶⁷ It may imply a less closely packed dimer and thus a larger RMSD value.

Interestingly, our calculations and other all-atom simulations also found T87–T87 contacts.^{60,68} This type of contact is due to the interhelix polar interactions between the hydroxyl groups of Thr87. In the concatenated result, $\sim 50\%$ of the time frames showed a distance between the hydroxyl groups of two Thr87 of less than 0.4 nm. Thr87 was in *g*– conformation in the initial simulated structure of GpA helices. In order to have side chain interaction between two Thr87, both Thr87 should rotate to adopt *g*+ conformation to bring their two hydroxyl groups closer. Our simulation results were consistent with the all-atom calculations which find that both Thr87 preferred *g*+ rather than *g*– conformation. As Thr was optimized to prefer *g*– conformation in our force field, a preference for *g*+ rotamer in this case is not an artifact of our model.³¹ This result suggests that interhelix H-bonds between Thr may play a role in stabilizing the GpA dimer.

CONCLUSION

The PACE force field has been extended to include lipids. The interactions between protein particles and lipid tails were parametrized by reproducing experimentally the free energy of solvation in hexadecane using small organic molecules. The average absolute error is about 1.0 kJ/mol. The transfer free energies from hexadecane to water were also calculated with an error of 1.1 kJ/mol. The interactions between protein particles and lipid heads were parametrized by fitting the corresponding PMFs obtained using all-atom simulations. Tilt angles of WALP and GWALP peptides and their mutants were calculated in different lipid bilayers. The results are in good agreement with the experimental data without having to make assumptions about the motions of the peptides. Association of glycophorin A helices was performed for 6 μ s. The simulated dimer in this work closely resembles the experimental structure. This is the first report showing the RMSD of the assembled GpA dimer lower than 0.3 nm compared to NMR structure. These results showcase the

high accuracy and efficacy of the extended PACE force field in studying membrane proteins.

■ ASSOCIATED CONTENT

S Supporting Information. Figure S1 shows schematic representation of hydrogen bonding potentials; Figure S2 shows interhelix separation of GpA helix dimer; Figure S3 shows contact map of the GpA dimer for the second and third runs; Table S1 outlines calculated tilt angles of WALP, GWALP peptides, and their mutants with original and modified backbone hydrogen-bonding potentials. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: chydwu@ust.hk.

■ ACKNOWLEDGMENT

We are grateful to the Research Grants Council of Hong Kong (663509) and Peking University for financial support of this research.

■ REFERENCES

- (1) Sachs, J. N.; Engelman, D. M. *Annu. Rev. Biochem.* **2006**, *75*, 707.
- (2) Wertén, P. J. L.; Rémy, H.-W.; de Groot, B. L.; Fotiadis, D.; Philippens, A.; Stahlberg, H.; Grubmüller, H.; Engel, A. *FEBS Lett.* **2002**, *529*, 65.
- (3) von Heijne, G. *Nat. Rev. Mol. Cell Biol.* **2006**, *7*, 909.
- (4) Phillips, R.; Ursell, T.; Wiggins, P.; Sens, P. *Nature* **2009**, *459*, 379.
- (5) Bowie, J. U. *Nature* **2005**, *438*, 581.
- (6) Roux, B.; Schulten, K. *Structure* **2004**, *12*, 1343.
- (7) Gumbart, J.; Wang, Y.; Aksimentiev, A.; Tajkhorshid, E.; Schulten, K. *Curr. Opin. Struct. Biol.* **2005**, *15*, 423.
- (8) Lindahl, E.; Sansom, M. S. P. *Curr. Opin. Struct. Biol.* **2008**, *18*, 425.
- (9) Ozdirekcan, S.; Etchebest, C.; Killian, J. A.; Fuchs, P. F. J. *J. Am. Chem. Soc.* **2007**, *129*, 15174.
- (10) Freddolino, P. L.; Liu, F.; Gruebele, M.; Schulten, K. *Biophys. J.* **2008**, *94*, L75.
- (11) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. *Science* **2010**, *330*, 341.
- (12) Shelley, J. C.; Shelley, M. Y.; Reeder, R. C.; Bandyopadhyay, S.; Klein, M. L. *J. Phys. Chem. B* **2001**, *105*, 4464.
- (13) Lopez, C. F.; Nielsen, S. O.; Srinivas, G.; DeGrado, W. F.; Klein, M. L. *J. Chem. Theory Comput.* **2006**, *2*, 649.
- (14) Venturoli, M.; Smit, B.; Sperotto, M. M. *Biophys. J.* **2005**, *88*, 1778.
- (15) de Meyer, F. J.-M.; Venturoli, M.; Smit, B. *Biophys. J.* **2008**, *95*, 1851.
- (16) Smeijers, A. F.; Pieterse, K.; Markvoort, A. J.; Hilbers, P. A. J. *J. Phys. Chem. B* **2006**, *110*, 13614.
- (17) Markvoort, A. J.; Smeijers, A. F.; Pieterse, K.; van Santen, R. A.; Hilbers, P. A. J. *J. Phys. Chem. B* **2007**, *111*, 5719.
- (18) Izvekov, S.; Voth, G. A. J. *J. Phys. Chem. B* **2009**, *113*, 4443.
- (19) Marrink, S. J.; de Vries, A. H.; Mark, A. E. J. *J. Phys. Chem. B* **2004**, *108*, 750.
- (20) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. *J. Phys. Chem. B* **2007**, *111*, 7812.
- (21) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S. J. *J. Chem. Theory Comput.* **2008**, *4*, 819.
- (22) Louhivuori, M.; Risselada, H. J.; van der Giessen, E.; Siewert, J.; Marrink, S. J. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 19856.
- (23) Smirnova, Y. G.; Marrink, S. J.; Lipowsky, R.; Knecht, V. *J. Am. Chem. Soc.* **2010**, *132*, 6710.
- (24) Shih, A. Y.; Arkhipov, A.; Freddolino, P. L.; Schulten, K. *J. Phys. Chem. B* **2006**, *110*, 3674.
- (25) Bond, P. J.; Sansom, M. S. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 2631.
- (26) Khalfa, A.; Treptow, W.; Maignet, B.; Tarek, M. *Chem. Phys.* **2009**, *358*, 161.
- (27) Shi, Q.; Izvekov, S.; Voth, G. A. J. *J. Phys. Chem. B* **2006**, *110*, 15045.
- (28) Rzepliela, A. J.; Louhivuori, M.; Peter, C.; Marrink, S. J. *J. Phys. Chem. Chem. Phys.* **2011**, *13*, 10437.
- (29) Han, W.; Wu, Y.-D. *J. Chem. Theory Comput.* **2007**, *3*, 2146.
- (30) Han, W.; Wan, C.-K.; Wu, Y.-D. *J. Chem. Theory Comput.* **2010**, *6*, 3390.
- (31) Han, W.; Wan, C.-K.; Jiang, F.; Wu, Y.-D. *J. Chem. Theory Comput.* **2010**, *6*, 3373.
- (32) Han, W.; Wan, C.-K.; Wu, Y.-D. *J. Chem. Theory Comput.* **2008**, *4*, 1891.
- (33) Jiang, F.; Han, W.; Wu, Y.-D. *J. Phys. Chem. B* **2010**, *114*, 5840.
- (34) van der Wel, P. C. A.; de Planque, M. R. R.; Greathouse, D. V.; Koeppe, R. E.; Killian, J. A. *Biophys. J.* **1998**, *74* (2), A304.
- (35) de Planque, M. R. R.; Bonev, B. B.; Demmers, J. A. A.; Greathouse, D. V.; Koeppe, R. E. II; Separovic, F.; Watts, A.; Killian, J. A. *Biochemistry* **2003**, *42*, 5341.
- (36) Kol, M. A.; van Laak, A. N. C.; Rijkers, D. T. S.; Killian, J. A.; de Kroon, A. I. P. M.; de Kruijff, B. *Biochemistry* **2003**, *42*, 231.
- (37) Weiss, T. M.; van der Wel, P. C. A.; Killian, J. A.; Koeppe, R. E. II; Huang, H. W. *Biophys. J.* **2003**, *84*, 379.
- (38) Strandberg, E.; Ozdirekcan, S.; Rijkers, D. T. S.; van der Wel, P. C. A.; Koeppe, R. E. II; Liskamp, R. M. J.; Killian, J. A. *Biophys. J.* **2004**, *86*, 3709.
- (39) Siegel, D. P.; Cherezov, V.; Greathouse, D. V.; Koeppe, R. E. II; Killian, J. A.; Caffrey, M. *Biophys. J.* **2006**, *90*, 200.
- (40) Sparr, E.; Ash, W. L.; Nazarov, P. V.; Rijkers, D. T. S.; Hemminga, M. A.; Tieleman, D. P.; Killian, J. A. *J. Biol. Chem.* **2005**, *280*, 39324.
- (41) Im, W.; Brooks, C. L., III *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6771.
- (42) Holt, A.; de Almeida, R. F. M.; Nyholm, T. K. M.; Loura, L. M. S.; Daily, A. E.; Staffhorst, R. W. H. M.; Rijkers, D. T. S.; Koeppe, R. E. II; Prieto, M.; Killian, J. A. *Biochemistry* **2008**, *47*, 2638.
- (43) Esteban-Martin, S.; Gimenez, D.; Fuentes, G.; Salgado, J. *Biochemistry* **2009**, *48*, 11441.
- (44) Vostrikov, V. V.; Daily, A. E.; Greathouse, D. V.; Koeppe, R. E., II *J. Biol. Chem.* **2010**, *285*, 31723.
- (45) Thomas, R.; Vostrikov, V. V.; Greathouse, D. V.; Koeppe, R. E., II *Biochemistry* **2009**, *48*, 11883.
- (46) Treutlein, H. R.; Lemmon, M. A.; Engelman, D. M.; Brunger, A. T. *Biochemistry* **1992**, *31*, 12726.
- (47) Langosch, D.; Brosig, B.; Kolmar, H.; Fritz, H. J. *J. Mol. Biol.* **1996**, *263*, 525.
- (48) MacKenzie, K. R.; Prestegard, J. H.; Engelman, D. M. *Science* **1997**, *276*, 131.
- (49) Brosig, B.; Langosch, D. *Protein Sci.* **1998**, *7*, 1052.
- (50) Fisher, L. E.; Engelman, D. M.; Sturgis, J. N. *J. Mol. Biol.* **1999**, *293*, 639.
- (51) Popot, J. L.; Engelman, D. M. *Annu. Rev. Biochem.* **2000**, *69*, 881.
- (52) Smith, S. O.; Song, D.; Shekar, S.; Groesbeek, M.; Ziliox, M.; Aimoto, S. *Biochemistry* **2001**, *40*, 6553.
- (53) Fisher, L. E.; Engelman, D. M.; Sturgis, J. N. *Biophys. J.* **2003**, *85*, 3097.
- (54) Petrache, H. I.; Grossfield, A.; MacKenzie, K. R.; Engelman, D. M.; Woolf, T. B. *J. Mol. Biol.* **2000**, *302*, 727.
- (55) Im, W.; Feig, M.; Brooks, C. L. *Biophys. J.* **2003**, *85*, 2900.

- (56) Kim, S.; Chamberlain, A. K.; Bowie, J. U. *J. Mol. Biol.* **2003**, *329*, 831.
- (57) Braun, R.; Engelman, D. M.; Schulten, K. *Biophys. J.* **2004**, *87*, 754.
- (58) Kokubo, H.; Okamoto, Y. *J. Chem. Phys.* **2004**, *120*, 10837.
- (59) Héning, J.; Pohorille, A.; Chipot, C. *J. Am. Chem. Soc.* **2005**, *127*, 8478.
- (60) Cuthbertson, J. M.; Bond, P. J.; Sansom, M. S. P. *Biochemistry* **2006**, *45*, 14298.
- (61) Efremov, R. G.; Vereshaga, Y. A.; Volynsky, P. E.; Nolde, D. E.; Arseniev, A. S. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 27.
- (62) Beevers, A. J.; Kukol, A. *J. Mol. Graphics Model* **2006**, *25*, 226.
- (63) Bond, P. J.; Sansom, M. S. P. *J. Am. Chem. Soc.* **2006**, *128*, 2697.
- (64) Elofsson, A.; Heijne, G. V. *Annu. Rev. Biochem.* **2007**, *76*, 125.
- (65) Metcalf, D. G.; Law, P. B.; DeGrado, W. F. *Proteins: Struct., Funct., Bioinf.* **2007**, *67*, 375.
- (66) Bu, L.; Im, W.; Brooks, C. L., III *Biophys. J.* **2007**, *92*, 854.
- (67) Psachoulia, E.; Fowler, P. W.; Bond, P. J.; Sansom, M. S. P. *Biochemistry* **2008**, *47*, 10503.
- (68) Psachoulia, E.; Marshall, D. P.; Sansom, M. S. P. *Acc. Chem. Res.* **2010**, *43*, 388.
- (69) Sengupta, D.; Marrink, S. J. *Phys. Chem. Chem. Phys.* **2010**, *12*, 12987.
- (70) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J. Comput. Chem.* **1997**, *18*, 1463.
- (71) Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. *Comput. Phys. Commun.* **1995**, *91*, 43.
- (72) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684.
- (73) Feenstra, K. A.; Hess, B.; Berendsen, H. J. C. *J. Comput. Chem.* **1999**, *20*, 786.
- (74) Mezei, M.; Beveridge, D. L. *Ann. N.Y. Acad. Sci.* **1986**, *482*, 1.
- (75) Faraldo-Gomez, J. D.; Smith, G. R.; Sansom, M. S. P. *Eur. Biophys. J.* **2002**, *31*, 217.
- (76) Kandt, C.; Ash, W. L.; Tieleman, D. P. *Methods* **2007**, *41*, 475.
- (77) Wolf, M. G.; Hoefling, M.; Aponte-Santamaria, C.; Grubmuller, H.; Groengof, G. *J. Comput. Chem.* **2010**, *31*, 2169.
- (78) Li, J.; Zhu, T.; Hawkins, G. D.; Winget, P.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. *Theor. Chim. Acta* **1999**, *103*, 9.
- (79) Abraham, M. H.; Whiting, G. S.; Fuchs, R.; Chambers, E. J. *J. Chem. Soc., Perkin Trans. 2* **1990**, 291.
- (80) Ozdirekcan, S.; Rijkers, D. T. S.; Liskamp, R. M. J.; Killian, J. A. *Biochemistry* **2005**, *44*, 1004.
- (81) Vostrikov, V. V.; Hall, B. A.; Greathouse, D. V.; Koeppe, R. E., II; Sansom, M. S. P. *J. Am. Chem. Soc.* **2010**, *132*, 5803.
- (82) MacCallum, J. L.; Bennett, W. F. D.; Tieleman, D. P. *Biophys. J.* **2008**, *94*, 3393.
- (83) MacCallum, J. L.; Bennett, W. F. D.; Tieleman, D. P. *J. Genet. Physiol.* **2007**, *129* (5), 371.
- (84) Kucerka, N.; Nieh, P. M.-P.; Katsaras, J. *Eur. Phys. J. E: Soft Matter Biol. Phys.* **2007**, *23*, 247.
- (85) Shirts, M. R.; Pitner, J. W.; Swope, W. C.; Pande, V. S. *J. Chem. Phys.* **2003**, *119* (11), 5740.
- (86) MacCallum, J. L.; Tieleman, D. P. *J. Comput. Chem.* **2003**, *24*, 1930.
- (87) Radzicka, A.; Wolfenden, R. *Biochemistry* **1988**, *27*, 1664.
- (88) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474.
- (89) Nymeyer, H.; Zhou, H.-X. *Biophys. J.* **2008**, *94*, 1185.
- (90) Yesylevskyy, S. O.; Schafer, L. V.; Sengupta, D.; Marrink, S. J. *PLoS Comput. Biol.* **2010**, *6*, e1000810.
- (91) Wang, J.; Kim, S.; Kovacs, F.; Cross, T. A. *Protein Sci.* **2001**, *10*, 2241.
- (92) Inbaraj, J. J.; Laryukhin, M.; Lorigan, G. A. *J. Am. Chem. Soc.* **2007**, *129*, 7710.
- (93) Kandasamy, S. K.; Larson, R. G. *Biophys. J.* **2006**, *90*, 2326.
- (94) Andrea, H.; Killian, J. A. *Eur. Biophys. J.* **2010**, *39*, 609.
- (95) de Planque, M. R. R.; Goormaghtigh, E.; Greathouse, D. V.; Koeppe, R. E.; Kruijtzter, J. A. W.; Liskamp, R. M. J.; de Kruijff, B.; Killian, J. A. *Biochemistry* **2001**, *40*, 5000.
- (96) Holt, A.; Koehorst, R. B.; Rutters-Meijneke, T.; Gelb, M. H.; Rijkers, D. T.; Hemminga, M. A.; Killian, J. A. *Biophys. J.* **2009**, *97*, 2258.
- (97) Esteban-Martin, S.; Salgado, J. *Biophys. J.* **2007**, *93*, 4278.
- (98) Strandberg, E.; Esteban-Martin, S.; Salgado, J.; Ulrich, A. S. *Biophys. J.* **2009**, *96*, 3223.
- (99) Im, W.; Lee, J.; Kim, T.; Rui, H. *J. Comput. Chem.* **2009**, *30*, 1662.
- (100) Kim, T.; Im, W. *Biophys. J.* **2010**, *99*, 175.
- (101) Bond, P. J.; Holyoake, J.; Ivetac, A.; Khalid, S.; Sansom, M. S. P. *J. Struct. Biol.* **2007**, *157*, 593.
- (102) Monticelli, L.; Tieleman, D. P.; Fuchs, P. F. *J. Biophys. J.* **2010**, *99*, 1455.
- (103) Holt, A.; Rougier, L.; Reat, V.; Jolibois, F.; Saurel, O.; Czaplicki, J.; Killian, J. A.; Milon, A. *Biophys. J.* **2010**, *98*, 1864.
- (104) Sankaramakrishnan, R.; Vishveshwara, S. *Int. J. Pept. Protein Res.* **1992**, *39*, 356.
- (105) Cordes, F. S.; Bright, J. N.; Sansom, M. S. P. *J. Mol. Biol.* **2002**, *323*, 951.
- (106) Lear, J. D.; Stouffer, A. L.; Gratkowski, H.; Nanda, V.; DeGrado, W. F. *Biophys. J.* **2004**, *87*, 3421.
- (107) Senes, A.; Engel, D. E.; DeGrado, W. F. *Curr. Opin. Struct. Biol.* **2004**, *14*, 465.
- (108) Popot, J. L.; Engelman, D. M. *Biochemistry* **1990**, *29*, 4031.
- (109) Russ, W. P.; Engelman, D. M. *J. Mol. Biol.* **2000**, *296*, 911.
- (110) Fleming, K. G. *J. Mol. Biol.* **2002**, *323*, 563.
- (111) Hong, H.; Blois, T. M.; Cao, Z.; Bowie, J. U. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 19802.
- (112) Senes, A.; Gerstein, M.; Engelman, D. M. *J. Mol. Biol.* **2000**, *296*, 921.
- (113) Lewis, B. A.; Engelman, D. M. *J. Mol. Biol.* **1983**, *166*, 211.

Characterization of the Ligand Receptor Encounter Complex and Its Potential for *in Silico* Kinetics-Based Drug Development

Karim M. ElSawy,^{*,†,‡} Reidun Twarock,^{†,‡,§} David P. Lane,^{||} Chandra S. Verma,[⊥] and Leo S. D. Caves^{†,‡}

[†]York Centre for Complex Systems Analysis (YCCSA), [‡]Department of Biology, and [§]Department of Mathematics, University of York, York YO10 5YW, United Kingdom

^{||}P53 Laboratory (p53Lab, A* STAR), 8A Biomedical Grove 06–06, Immunos, Singapore 138648

[⊥]Bioinformatics Institute (A*STAR), 30 Biopolis Str., 07–01 Matrix, Singapore 138671

 Supporting Information

ABSTRACT: The study of drug–receptor interactions has largely been framed in terms of the equilibrium thermodynamic binding affinity, an *in vitro* measure of the stability of the drug–receptor complex that is commonly used as a proxy measure of *in vivo* biological activity. In response to the growing realization of the importance of binding kinetics to *in vivo* drug activity we present a computational methodology for the kinetic characterization of drug–receptor interactions in terms of the encounter complex. Using trajectory data from multiple Brownian dynamics simulations of ligand diffusion, we derive the spatial density of the ligand around the receptor and show how it can be quantitatively partitioned into different basins of attraction. Numerical integration of the ligand densities within the basins can be used to estimate the residence time of the ligand within these diffusive binding sites. Simulations of two structurally similar inhibitors of Hsp90 exhibit diffusive binding sites with similar spatial structure but with different ligand residence times. In contrast, a pair of structurally dissimilar inhibitors of MDM2, a peptide and a small molecule, exhibit spatially distinct basins of attraction around the receptor, which in turn reveal differences in ligand orientational order. Thus, our kinetic approach provides microscopic details of drug–receptor dynamics that provide novel insight into the observed differences in the thermodynamic binding affinities for the two inhibitors, such as the differences in the entropic contributions to binding. The characterization of the encounter complex, in terms of the structure, topology, and dynamics of diffusive binding sites, offers a new perspective on ligand–receptor interactions and the potential for greater insight into drug action. The method, which requires no prior knowledge of the bound state, is a first step toward the incorporation of ligand kinetics into *in silico* drug development protocols.

INTRODUCTION

The *status quo* of drug development is the selective and optimal targeting of low-molecular-weight compounds to a particular bioactive molecule.¹ In general, this is carried out by *in vitro* optimization of the drug–target association constant (K_a), free energy of association (ΔG_a), or the half-maximal inhibitory concentration (IC_{50}). In this context, these parameters are used as quantitative measures (or proxies) of the drug biological activity.² This approach implicitly assumes closed-system conditions in which the target is exposed to an invariant concentration of the drug and thermodynamic equilibrium is assumed.³ However, *in vivo* conditions are very different, as drug concentration is no longer invariant due to factors such as circulation, absorption, metabolism, and interaction with other cellular constituents. Under these conditions, it is becoming increasingly clear that the kinetics of drug binding, as measured by the association and dissociation rate constants (k_{on} and k_{off}) of drug target interactions, are of significant relevance to its biological activity.^{1,4}

Initial work in developing quantitative structure–property relationships (QSPR) based on binding kinetics has confirmed the utility of both k_{on} and k_{off} as important and distinct factors relating to pharmacological and pharmacokinetic properties.⁵ It has been observed that, in lead optimization, a decrease in the association constant (K_d) often results in a lowering of k_{off} by

proxy, due to the typically smaller dynamic range of k_{on} . Unlike k_{off} , k_{on} is usually diffusion-controlled and hence not amenable to optimization. Indeed, increasing k_{off} was found to be strongly related to HIV-1 protease resistance to the inhibitor saquinavir.⁶ This was further confirmed by a study of the kinetics of the interaction between drug-resistant variants of HIV-1 protease and several clinically used inhibitors (amprenavir, indinavir, nelfinavir, ritonavir, and saquinavir), where the reduction of affinity was related to a combination of decreased k_{on} and increased k_{off} .⁷ Similarly, the development of resistance in EGFR has been associated with altered kinetics,⁸ where p53 DNA site-specific recognition was found to rely more on differences in binding k_{off} rather than on differences in affinities.⁹ These observations have led a number of recent researchers to stress a greater role for drug binding kinetics in therapeutic differentiation strategies,¹⁰ mitigating off-target mediated toxicity and leading to improved drug safety and tolerability.⁴ It is becoming clear that the off-rate k_{off} or the drug–target residence time τ ($\tau \sim 1/k_{off}$) is an important and perhaps crucial property for drug lead optimization,^{1,4,11–13} thereby providing an alternative route to improving the therapeutic utility of a drug.¹⁰

Received: August 11, 2011

Published: December 28, 2011

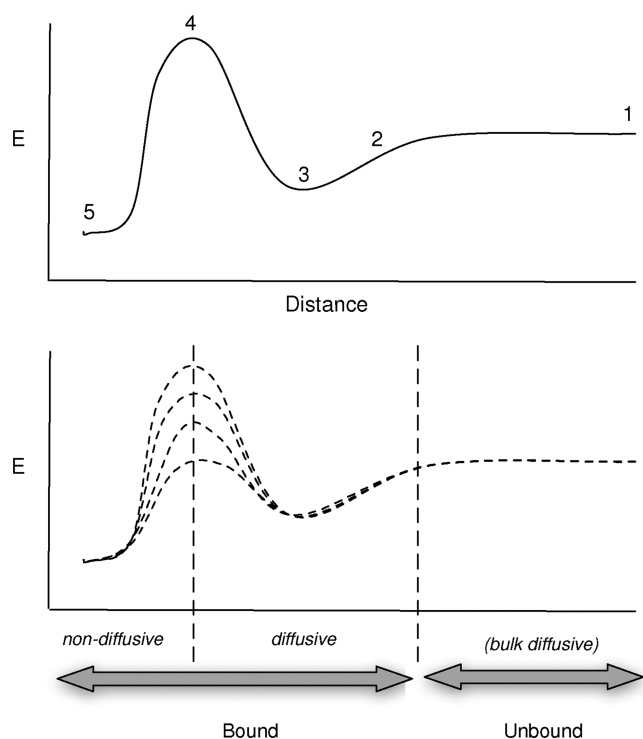


Figure 1. (Upper) Traditional view of the free energy profile along the reaction coordinate of ligand–receptor interaction. Different domains of interaction are schematically depicted: (1) free diffusion, (2) electrostatic steering, (3) encounter complex, (4) energy barrier crossing, and (5) formation of the bound (final) complex. (Lower) An alternative view with two broad regimes of interaction: unbound (free bulk diffusion) and bound, which incorporates both the restricted diffusion of the encounter complex and the nondiffusive characteristics of (final) bound states. In the general case, due to ligand and receptor conformational flexibility, the energy surface is dynamic—and with the variation in 4 shown for illustration. Under nonequilibrium conditions, both the occupation of states and the barrier height are time-dependent.

We consider key stages of ligand–receptor interactions via an illustrative reaction coordinate (Figure 1). In this scheme, the traditional focus is on the energy barrier (labeled 4) separating the freely diffusing ligand (labeled 1) from the bound complex (labeled 5). At equilibrium, this barrier can be related to the association and dissociation rates (k_{on} and k_{off}). However, prior to reaching equilibrium, an encounter complex is formed. The encounter complex corresponds to the configuration of the ligand receptor complex prior to crossing the energy barrier to the final bound state. The process of ligand–receptor interaction can, therefore, be described in terms of three distinct regimes: *unbound* (in bulk solution), *nondiffusively bound* (the classic binding site, and the focus of the thermodynamic approach to drug interaction), and an intermediate *diffusively bound* regime characterized by the *encounter complex* (Figure 1). The ligand in this diffusively bound regime is spatially restrained in the vicinity of the receptor due to net associative influx.

The encounter complex is therefore a prerequisite for the binding reaction to proceed and is differentiated from other configurations by the requirement that *the ligand association rate is much higher than its dissociation rate*.¹⁴ The residency of the encounter complex increases the possibility of mutual perturbation of the ligand and the receptor conformational landscape, which may lead to a reduction of the barrier height. This is consistent

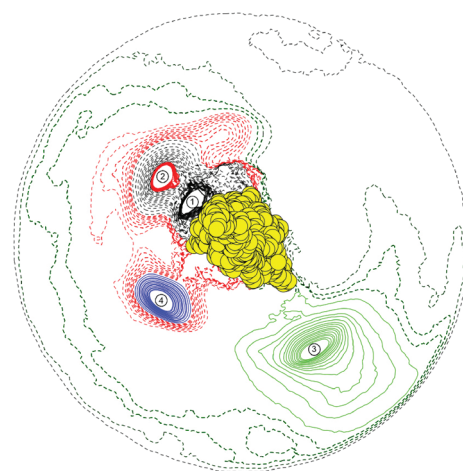


Figure 2. A 2D illustration of the partitioning of the space around the MDM2 receptor (yellow) into four different basins of attraction (shown as solid lines) and their connecting superbases (shown as dashed lines). The basins of attraction correspond to regions of space where the rate of association of the ligand–receptor encounters is higher than the rate of dissociation–formation of the ligand–receptor encounter complex. This is ensured by two constraints: (1) the inward direction of the gradient of the ligand density contours in these regions (i.e., inward overall ligand influx) and (2) the requirement that the contours be closed such that the ligand motion in these region is spatially restricted. Hot spots within these basins are labeled in ascending order according to the corresponding ligand residence time.

with the observation that the forward rate constant k_{on} is diffusion-limited in a large number of biological processes.^{15–17}

Amidst the flux and clearance that is characteristic of the *in vivo* situation, the encounter complex represents a higher local availability (or “concentration”) of the drug around the target. This sustains the influence of the drug on the target (i.e., perturbation of receptor conformational states, via conformational selection, or induced fit), which in turn is responsible for the biological response.¹ Thus, the residence time of the encounter complex emerges as an important, and as yet underexplored, kinetic factor in ligand–receptor interactions in a pharmacological context.

Currently, myriad computational tools exist to support drug development within the thermodynamic approach.^{18–21} More recently, steered molecular dynamics simulations have enabled the successful discrimination of good and bad binders among a set of molecules to a receptor, thus pointing toward the relevance of k_{off} .²² In addition, with the availability of computers that enable much longer simulation times, atomistic simulations have revealed the pathways of drug binding to receptors and have displayed apparent kinetics that are close to the experimentally determined ones.^{23,24} However, both of these methods require very long simulations that are currently not generally accessible, which in turn make them not amenable to easy statistical analyses. Moreover, they do not provide a general framework for addressing the kinetics of ligand protein interactions.

Here, we introduce a novel method to compute kinetic information related to drug–target interactions that could be a powerful and complementary addition to the drug development toolkit.²⁵ We describe a computational strategy for the characterization of the ligand–receptor *encounter complex*. On the basis of the definition introduced above, we regard the encounter complex as being in terms of *basins of attraction* of the ligand

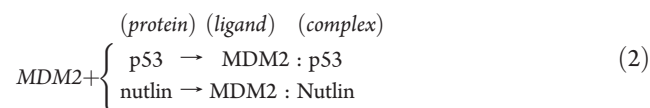
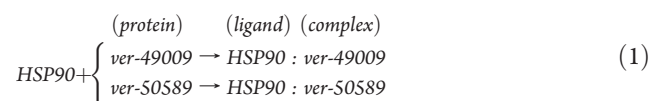
around the receptor, and we provide a method for their identification. We estimate the residence time in these basins of attraction by integration of the ligand spatial probability density derived from multiple ligand trajectories in Brownian dynamics simulations.

The use of a clear physical criterion in defining the encounter complex allows for a blind (i.e., unbiased) characterization of multiple basins of attraction representing kinetically distinct diffusively bound states with their own characteristic residence times (Figure 2). This framework offers the possibility of significantly supplementing the current models of drug-receptor interactions with quantitative parameters relating to the structure (spatial distributions of the drug within the basins of attraction), topology (connectivity of the basins), and dynamics (flux in, out, and between basins) within the diffusive-bound regime.

MATERIAL AND METHODS

Diffusional Dynamics of Ligand Receptor Interaction.

Brownian dynamics (BD) simulations of the ligand protein diffusional encounters were carried out for the association schemes 1 and 2 using the SDA package.^{26,27}



Brownian Dynamics Setup. The BD trajectories were propagated by solving the translational and rotational diffusion equations, using the Ermak–McCammon algorithm²⁸ as implemented in the SDA package version 4.23b. The diffusion coefficients were calculated using the HYDROPRO software.²⁹ Initially, the center of mass (COM) of the protein was placed at the origin, and the ligand was placed at a $b = 150.0 \text{ \AA}$ COM–COM separation relative to the protein center of mass. At this separation, there is no preferential orientation of the ligand since the electrostatic potential of the protein is nearly isotropic at distances greater than 80 \AA from its center of mass. A time step of 0.1 ps was used when the COM–COM separation was less than 90 \AA . At larger separations, the time step was increased linearly with a slope of 0.5 ps \AA^{-1} . The simulations were terminated if the ligand–protein COM–COM separation exceeded $c = 3b \text{ \AA}$.

In scheme 1, multiple conformations of the VER-49009 and VER-50589 ligands were used as initial structures for BD simulation of the ligand diffusion around the HSP90 protein (PDB ID: 2uwd). The ligand conformations were selected on the basis of principal component analysis (PCA) of constant-temperature MD trajectories (1 ns) of the unbound ligand using the Merck force field within CHARMM. The selected ligand conformations correspond to high-density regions in the ligand conformational landscape as revealed by PCA. In scheme 2, 15 snapshots of the ligand protein conformations were extracted from equilibrated MD trajectories of the p53 peptide (ETFSDLWKLLEN) and nutlin-2 MDM2 bound complexes.³⁰ In both schemes, a large number of BD trajectories (100 000 and 750 000 for schemes 1 and 2, respectively) were generated and checked for convergence with respect to the total residence time of the ligand around

the receptor (see below). Typically, 50 000 BD trajectories were found to be sufficient to reach convergence in both schemes.

In the BD simulations, the forces between the ligand and the protein derive from steric, desolvation, and electrostatic interactions. Steric interactions were accounted for implicitly through the use of steric exclusion grids (grid spacing of 1 \AA) centered on both protein and ligand. The electrostatic force on any atom of the ligand was calculated by multiplying its charge by the electrostatic potential of the protein at that spatial position. The electrostatic potential around the protein and ligand was calculated by numerically solving the nonlinear Poisson–Boltzmann equation^{31,32} on a grid with dimensions of $161 \times 161 \times 161 \text{ \AA}$ centered at the ligand/protein using the APBS program.³³ Using a grid spacing of 1 \AA (as required by the SDA package) results in computational efficiencies in the BD simulations. Adapted PEOE atomic charges and radii of the ligand and the protein atoms were assigned using the PDB2PQR program.^{34,35} The solvent dielectric constant was set to 78.5 and the protein interior dielectric constant to 4; the salt concentration was set to 0.15 M . The solute–solvent boundary was defined at the van der Waals surface because molecular surface definition was found to result in significant underestimation of the association rates in some cases.³⁶ During the BD simulations, for the sake of computational efficiency, the full set of atomic charges of the ligand was replaced by a smaller set of effective charges that accurately reproduced their calculated electrostatic potential.³⁷ The effective charges were derived by the ECM module in the SDA package so as to reproduce the electrostatic potential at the accessible surface (defined by a probe of 4 \AA) in a 3-\AA -thick layer extending outward from each structure.

Desolvation effects were incorporated through a desolvation penalty grid³⁸ around the ligand and the protein using a scaling factor of 1.67, consistent with the surface definition for the solute–solvent boundary.³⁶ Use of explicit water has been shown in a number of studies to be an important factor in molecular association^{39,40} and the formation of the ligand-bound complex⁴¹ (labeled 5; Figure 1). However, since we focus on the encounter complex (labeled 3; Figure 1), outside its conventional binding site, we believe, an implicit account of desolvation effects using a penalty grid is a reasonable choice consistent with the use of BD simulations.

Convergence of the BD Simulation. In order to check for convergence of the BD simulations, the total residence time of the ligand around the protein was estimated by aggregating successively larger numbers of BD trajectories in post processing. The total residence time of the ligand was then computed via integration of the residence time radial profile. Since the BD trajectories are independent, there is no particular order in which to aggregate the data. To provide a robust estimate of convergence, the following procedure was used. The BD trajectories were split randomly into bins that comprise 5000 trajectories each; the bins were then sampled without replacement and spliced randomly to form bigger bins that comprise 10 000, 15 000, 20 000, etc. trajectories. The process was repeated 1000 times, and the average residence time of the ligand around the protein was then computed for each bin (Supporting Information, Figure 1). The BD simulations were found to converge beyond 50 000 trajectories for both schemes 1 and 2.

Kinetics-Based Approach for the Identification of Encounter Complex Basins of Attraction. The following algorithm forms the basis of our kinetics-based approach to identifying the dynamically derived ligand spatial density around the receptor and partitioning it into distinct basins of attraction, which comprise

the ligand–receptor encounter complex. From the trajectory data, a 3D spatial probability density grid is constructed around the receptor by computing the average frequency of the ligand center of mass visiting individual spatial grid cells. The grid dimensions were chosen so as to extend 200 Å in each direction with a grid spacing of 1 Å. The root-mean-square displacement of the center of mass of the ligand during the BD simulations (~ 0.15 Å in schemes 1 and 2) sets a lower limit for the grid spacing that can be used. The choice of a grid spacing of 1 Å reflects a balance between retaining dynamical information and tractable spatial resolution: in practice, it generates a detailed and smooth spatial probability density landscape.

The 3D ligand spatial density grid is contoured at equally spaced contour intervals that extend from the highest density (strongly interacting with the receptor: diffusely bound regime) to the lowest density (weak/noninteracting: diffusive regime). Closed contours at the highest level of density represent interaction “hot spots”. Closed contour surfaces at successively lower density levels that encompass each individual hot spot are then accumulated. As a consequence of following the density gradient in this manner, the ligand flux in these regions is such that the average rate of inward ligand association (i.e., toward the hot spot) is higher than its average outward dissociation rate. As the contour surfaces are “closed”, the process therefore ensures that the direction of the ligand density gradient is maintained inward from all directions. These characteristics, which result from restriction (or channelling) of the ligand motion in the diffusional trajectories, result in these regions acting as basins of attraction. The connectivity between different basins of attraction is determined by detecting closed contour surfaces that encompass multiple hot spots: such regions represent superbasins of attraction within a global disconnectivity tree.⁴²

This scheme, based on ligand–receptor binding dynamics, yields a partitioning of the ligand spatial density around the receptor into kinetically distinct, spatially resolved encounter complex basins of attraction that correspond to putative binding sites in the diffusively bound regime. The identification of these sites does not require prior knowledge of a (conventional nondiffusive) binding site on the receptor, or the use of other *ad hoc* criteria.^{14,27,43} This is notable, in view of the increasing importance of the role of allosteric sites in modulating activity⁴⁴ which are difficult to infer from static methods such as X-ray crystallography.

Estimation of Ligand Site-Specific Residence Time (τ_{LR}). The ligand *site-specific residence time* is computed for individual basins of attraction by numerical integration of the time-averaged ligand probability density in each basin of attraction. Using small contour intervals, in which the ligand density D is effectively constant, the integral can be approximated by

$$\tau_{LR} = \sum_{n=2}^N \Delta t (V_n - V_{n-1}) (D_n + D_{n-1}) / 2$$

where $(V_n - V_{n-1})$ is the volume enclosed between two consecutive contour surfaces at ligand densities D_n and D_{n-1} and Δt is the time step used in the BD simulation. In the case of multiple basins of attraction, it may be convenient to express the *fractional residence time* of a given basin as a proportion of the total residence time over all basins.

RESULTS AND DISCUSSION

In order to illustrate the applicability of our approach, we considered two ligand–receptor systems: (1) interactions of two

structurally homologous inhibitors (VER-50589 and VER-49009) with their target Hsp90 and (2) interactions of two very different ligands, an α -helical peptide mimic of p53 and a nonpeptide ligand (Nutlin), with MDM2, the negative regulator of p53.

1. Hsp90–Inhibitor Interactions. Hsp90 inhibitors cause the inactivation and eventual degradation of Hsp90 client proteins and have shown promising antitumor activity in preclinical model systems.⁴⁵ However, the relationship between the biological activity and thermodynamic characteristics of a number of these inhibitors remains unclear.⁴⁶ Recently, a potent isoxazole analogue of 3,4-diarylpyrazole resorcinols, called VER-50589, has been identified. This compound is 9-fold more potent than a highly homologous analogue, the pyrazole VER-49009.⁴⁶ Interestingly, the key difference between the two compounds is the introduction of an amine (N–H) group (and thus a new H-bond donor) into a heterocyclic ring, resulting from the replacement of an oxygen atom in VER-50589 by a nitrogen atom in VER-49009 (see Figure 3a, top).

The crystal structures of these inhibitors bound to the N-terminal domain of the Hsp90 monomer reveal a “virtually identical binding mode”;⁴⁶ calorimetric analysis revealed a tighter binding for VER-50589 vs VER-49009 (K_d : 4.5 vs 78.0 nmol L⁻¹) attributed to a higher enthalpy of VER-50589 binding, yet the rationale for this remains unclear.⁴⁶ However, a kinetic analysis of ligand binding has suggested that the higher cellular activity of VER-50589 relative to VER-49009 may be associated with an approximately 10-fold slower off-rate for VER-50589 compared to VER-49009, leading to higher cellular concentrations.⁴⁷

BD simulations (Figure 3) reveal that both inhibitors exhibit very similar encounter complex distributions around the Hsp90 protein (Figure 3b) forming three main basins of attraction in proximity to the residues Thr:65, Gly:125, and Glu:205 on the receptor surface (VER-50589 exhibits a very minor fourth basin). A comparison of the fractional residence times of the ligands within these basins shows two distinct patterns: VER-49009 exhibits an essentially equal distribution of the ligand across the three basins (fractional residence times of ~ 35 , 35, and 29%), whereas VER-50589 preferentially occupies a single basin (fractional residence times of 50, 33.5, 14%). The unequal occupancy within the three basins leads to differences in the effective local concentration of the ligand.

Although we cannot draw firm conclusions from these preliminary results, we can begin to illustrate how this new picture of drug–target interactions could provide additional insight into the mechanism of drug action. Without consideration of major allosteric effects, a working assumption is that the degree of influence that a ligand has on the receptor is proportional to the amount of time it is available to exert that influence. In this context, the encounter complex basin residence time profile of VER-49009 suggests a higher degree of ligand depletion than VER-50589. By contrast, VER-50589 spends more time residing in a localized region close to the receptor. Such characteristics are of interest in relation to the superior pharmacodynamics (longer duration of action) and enhanced antitumor efficacy of VER-50589 vs VER-49009 *in vivo*⁴⁶ and may offer new perspectives on future drug development.

2. MDM2–Inhibitor Interactions. The protein p53 protects cells from various sorts of damage, and its levels are controlled in a negative feedback loop with the ligase MDM2.⁴⁸ This interaction is critical for therapeutic intervention in tumors that have overexpressed MDM2, leading to an intense search for MDM2 inhibitors.^{49,50} There are two major avenues being explored for

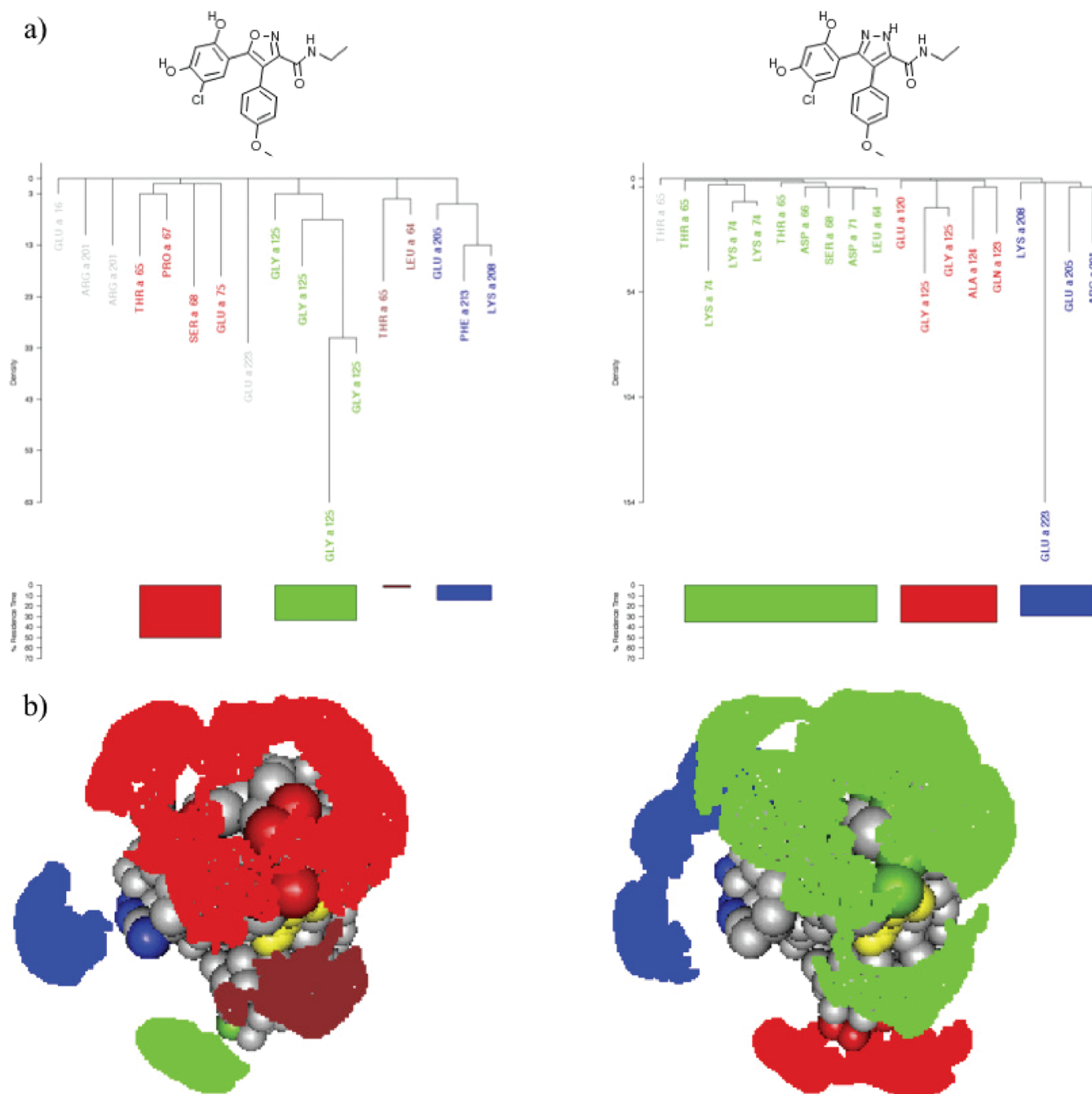


Figure 3. (a) The disconnectivity trees of the density of the ligand–Hsp90 encounter complex and the corresponding fractional residence times of the different basins of attraction of VER-50589 (left) and VER-49009 (right) upon interaction with the Hsp90 N-terminal domain. The leaves of the disconnectivity tree correspond to different basins of attraction; sets of connected basins form superbasins that are shown in different colors such that those corresponding to the superbasin with the highest residence time are shown in red. The distribution of the center of mass of the ligand within these superbasins around the Hsp90 is shown in b using the same color scheme. Residues closest to the basins within each superbasin are labeled accordingly in a and b. The ligand within the active site as determined by X-ray crystallography (PDB ID: 2uwd) is shown in yellow for each case.

inhibitors: small molecules (e.g., Nutlins) and peptide mimics of p53, both of which disrupt or prevent p53–MDM2 interactions; indeed some are now entering clinical trials. However, in general, the peptidic inhibitors have not been as effective as Nutlins,⁵¹ a trend that is in line with their relative thermodynamic binding affinity.⁵² Here, we apply our methodology to examine how binding kinetics relate to these observed differences in activity. In contrast to the Hsp90 inhibitors, the distribution of the basins of

attraction of Nutlin and the α -helical peptide around MDM2 are distinctly different (Figure 4a). This reveals that the characteristics of interactions with a given receptor (and thus potentially the mechanism of action) can vary between ligands. This is important, as when developing potential inhibitors, such as nutlins and peptidomimetics, ligand optimization is currently carried out in terms of thermodynamic parameters, assuming a largely invariant binding site and thus similar modes of interactions. In contrast,

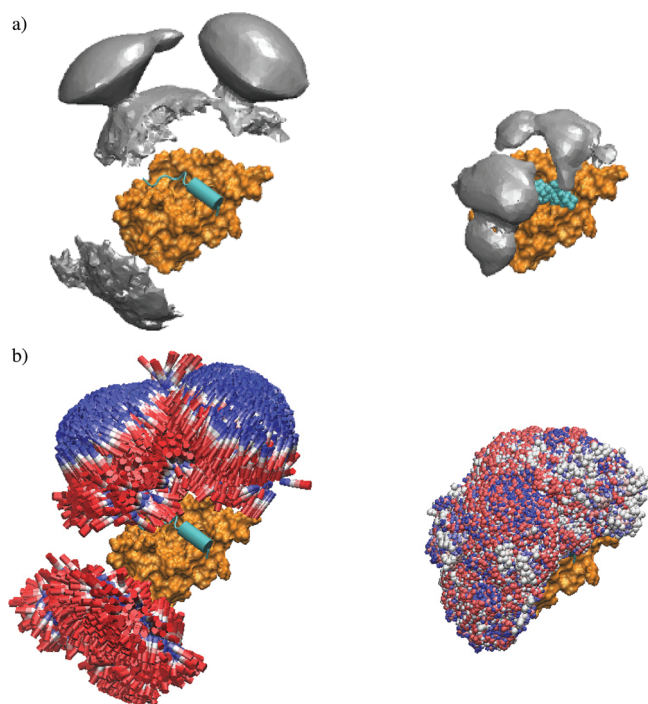


Figure 4. (a) The basins of attraction (gray) of the α -helical peptide ligand (left) and Nutlin-2 (right) upon interaction with the MDM2 (orange). The final bound complex is shown in cyan in cartoon and VDW representations. The corresponding structures of the encounter complex within each basin are shown in b. In order to illustrate the direction of the molecular structures within these basins, the structures are colored according to the atom index of each of them using a color palette that changes smoothly from blue (first atom) to red (last atom).

the kinetic profiles, in terms of encounter complex basins of attraction, suggest that this assumption needs to be re-examined. In agreement with the discovery of the importance of allosteric interactions in peptide-MDM2 interactions,³⁰ the spatial distribution of the basins of attraction of the α -helical peptide suggests that the initial diffusive encounter between the peptide and MDM2 takes place far from the (classic nondiffusive) binding site, at the N- and C-termini of MDM2 (Figure 4a, left). Moreover, within the basins of attraction, the distribution of ligand configurations exhibits a relatively high degree of order in terms of helix orientation (Figure 4b, left). This reflects the significant degree of loss of rotational entropy upon peptide interaction with the MDM2 that has to be compensated for by enthalpic interactions upon formation of the final bound complex.⁵² The orientational order of the peptides clearly stems from alignment of the ligand helix dipole moment with the electrostatic potential of the MDM2 receptor. Interestingly, the peptide helix orientation with respect to the receptor surface is reversed within the two basins. By comparison, interaction of Nutlin with MDM2 takes place via two basins of attraction that directly interact with the clefts of the (classic nondiffusive) binding site on the MDM2 surface (Figure 4a, right). By contrast, the distribution of structures of the Nutlin encounter complex within these basins does not exhibit the same degree of order as observed for the peptide inhibitor (Figure 4b, right), suggesting that as a consequence of the orientational preordering, the entropic loss upon Nutlin binding to MDM2 will be smaller than the peptide, thus contributing to its higher affinity.

This new kinetics-based mechanistic insight complements the traditional thermodynamic approach to drug optimization, which focuses attention on the (nondiffusive) bound state, which necessarily has a more restricted configurational distribution than diffusively bound encounter complex states. In the thermodynamic approach, optimization (in terms binding affinity) of peptidic ligands would be based on maximizing the enthalpic contribution to binding in order to compensate for the entropic loss on moving from the (bulk diffusive) unbound regime. From our kinetics-based analysis of the encounter complex structural distribution, we see the role of the helix dipole in restricting the orientational freedom of the peptide ligand, compared to Nutlin. This observation could provide new avenues for exploitation in drug optimization, for example, to tune binding properties by engineering peptidic ligand dipole moments to modify the configurational ordering in the encounter complex—a new approach that emerges from our methodology.

CONCLUSIONS

We present an effective and principled computational approach for the kinetic characterisation of ligand–receptor interactions with a focus on the encounter complex, defined clearly as states where the ligand–receptor association rate is higher than the dissociation rate. Our approach to identifying these diffusive binding sites does not rely upon the assumption of reaching thermodynamic equilibrium or on prior knowledge of the bound state. Instead, we adopt a dynamical approach: generating trajectories representing ligand–receptor diffusional encounters on a natural time scale (subject to certain technical constraints), representing the time for the ligand to diffusively engage and subsequently disengage from the receptor. Using data from multiple trajectories, we derive spatial densities of the ligand around the receptor. The ligand spatial distribution can then be partitioned into distinct basins of attraction, representing diffusive binding sites. Residence times within these sites are estimated by numerical integration of the corresponding ligand densities. We emphasize that the residence times estimated in this procedure relate to the encounter complex and do not incorporate contributions arising from the classic bound state, which is the subject of ongoing work.

The importance of the residence time in shaping the biological activity of drug candidates has been the focus of several recent studies.^{1,3,4,12} Thus, a quantitative structural and kinetic characterization of diffusively bound ligand–receptor interactions provides additional information that may provide greater (or at least complementary) insight into the “mechanism” of drug action *in vivo* than that provided by traditional thermodynamic approaches. Specifically, the introduction of a new set of microstates (encounter complex basins of attraction) opens up new avenues for gaining a greater understanding of the relationship of the microscopic ligand–receptor dynamics to the effective macrostates that underpin thermodynamic treatments of the binding process.

We have applied the methodology to two different systems: Hsp90 inhibitors and MDM2 inhibitors. For Hsp90, we found that the redistribution of the ligand residence time across the different basins of attraction is likely to be an important factor in shaping the biological activity of structurally related inhibitors. For MDM2, on the other hand, we showed that structurally unrelated inhibitors, an α -helical peptide and Nutlin, exhibit two distinctly different patterns of interaction with the MDM2 receptor. The basins of attraction of the α -helical peptide around MDM2 are located far from the ligand binding site, suggestive of

an allosteric mechanism, while the locations of the basins for Nutlin are indicative of more direct interactions with the classic ligand binding site. In our study, the peptide conformation is restricted to be largely α -helical. This is justified by reference to the increasing use of “stapled” peptides as therapeutic agents which rely on stabilizing the peptide into an α -helical conformation through the use of a hydrocarbon bridge.^{53–55} The introduction of the bridge does not significantly affect the electrostatic potential of the peptide, and therefore we anticipate that its absence in our model will not significantly affect our results since electrostatic interaction is the main driving force for the formation of the encounter complex.

In this proof of principle study, we employ multiple static conformations of the ligand and receptor and thus cannot address cooperative features of molecular recognition such as induced fit of receptor–ligand conformational states.⁵⁶ However, such effects can be incorporated into the approach, albeit at additional computational cost, by the use of nested BD/MD simulations, which are the subject of ongoing study.

The methodology is timely given the growing realization of the importance of drug binding kinetics, particularly the drug–target residence time, for optimizing drug efficacy *in vivo*. The additional insight provided by this kinetics-based characterization could provide a powerful complement to the traditional structure-based computational drug optimization techniques that are largely focused on the thermodynamics of drug–receptor interactions in the classic (nondiffusive) binding site. Of particular promise is the potential for additional insight into the mechanisms of drug action afforded by a richer picture of molecular interactions with the adoption of a complex dynamical (or even adaptive) systems framework (e.g., trajectories, basins of attraction), over traditional equilibrium thinking (e.g., states, static distributions). We offer this approach as an example of a new breed of computational tools to explore the new paradigm of kinetic-based drug development.

■ ASSOCIATED CONTENT

S Supporting Information. Convergence of the residence time as a function of the number of BD trajectories. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: karim.elsawy.email@gmail.com.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENT

We thank Prof. Roderick E. Hubbard for many helpful discussions and for suggesting the Hsp90 test case. We are very grateful to Drs. Garib Murshudov and Seishi Shimizu for the provision of computing resources.

■ REFERENCES

- (1) Copeland, R. A.; Pompliano, D. L.; Meek, T. D. Drug–target residence time and its implications for lead optimization. *Nat. Rev. Drug Discovery* **2006**, *5* (9), 730–739.
- (2) Kenakin, T. Quantifying Biological Activity in Chemical Terms: A Pharmacology Primer to Describe Drug Effect. *ACS Chem. Biol.* **2009**, *4* (4), 249–260.

- (3) Tummino, P. J.; Copeland, R. A. Residence time of receptor–ligand complexes and its effect on biological function. *Biochemistry* **2008**, *47* (20), 5481–92.

- (4) Copeland, R. A. The dynamics of drug–target interactions: drug–target residence time and its impact on efficacy and safety. *Expert Opin. Drug Discovery* **2010**, *5* (4), 305–310.

- (5) Andersson, K.; Hämäläinen, M. D. Replacing affinity with binding kinetics in QSAR studies resolves otherwise confounded effects. *J. Chemom.* **2006**, *20* (8–10), 370–375.

- (6) Maschera, B.; Darby, G.; PalÅ°, G.; Wright, L. L.; Tisdale, M.; Myers, R.; Blair, E. D.; Furfine, E. S. Human Immunodeficiency Virus. Mutations in the viral protease that confer resistance to saquinavir increase the dissociation rate constant of the protease–saquinavir complex. *J. Biol. Chem.* **1996**, *271* (52), 33231–33235.

- (7) Shuman, C. F.; Markgren, P. O.; Hamalainen, M.; Danielson, U. H. Elucidation of HIV-1 protease resistance by characterization of interaction kinetics between inhibitors and enzyme variants. *Antiviral Res.* **2003**, *58* (3), 235–42.

- (8) Yun, C.-H.; Mengwasser, K. E.; Toms, A. V.; Woo, M. S.; Greulich, H.; Wong, K.-K.; Meyerson, M.; Eck, M. J. The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105* (6), 2070–2075.

- (9) Petty, T. J.; Emamzadah, S.; Costantino, L.; Petkova, I.; Stavridi, E. S.; Saven, J. G.; Vauthey, E.; Halazonetis, T. D. An induced fit mechanism regulates p53 DNA binding kinetics to confer sequence specificity. *EMBO J.* **2011**, *30* (11), 2167–2176.

- (10) Swinney, D. C. Applications of Binding Kinetics to Drug Discovery: Translation of Binding Mechanisms to Clinically Differentiated Therapeutic Responses. *Pharm. Med.* **2008**, *22* (1), 23–34.

- (11) Swinney, D. C. The role of binding kinetics in therapeutically useful drug action. *Curr. Opin. Drug Discovery Dev.* **2009**, *12* (1), 31–9.

- (12) Zhang, R.; Monsma, F. The importance of drug–target residence time. *Curr. Opin. Drug Discovery Dev.* **2009**, *12* (4), 488–96.

- (13) Lu, H.; Tonge, P. J. Drug–target residence time: critical information for lead optimization. *Curr. Opin. Chem. Biol.* **2010**, *14* (4), 467–74.

- (14) Gabdouliline, R. R.; Wade, R. C. On the protein–protein diffusional encounter complex. *J. Mol. Recognit.* **1999**, *12* (4), 226–34.

- (15) Tzafiriri, A. R.; Levin, A. D.; Edelman, E. R. Diffusion-limited binding explains binary dose response for local arterial and tumour drug delivery. *Cell Proliferation* **2009**, *42* (3), 348–363.

- (16) Raman, C. S.; Jemmerson, R.; Nall, B. T.; Allen, M. J. Diffusion-limited rates for monoclonal antibody binding to cytochrome c. *Biochemistry* **1992**, *31* (42), 10370–10379.

- (17) McGahay, V. Inertial effects and diffusion. *J. Non-Cryst. Solids* **2004**, *349*, 234–241.

- (18) Zoete, V.; Grosdidier, A.; Michielin, O. Docking, virtual high throughput screening and *in silico* fragment-based drug design. *J. Cell. Mol. Med.* **2009**, *13* (2), 238–248.

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

- (20) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophys. J.* **1997**, *72* (3), 1047–1069.

- (21) Durrant, J. D.; McCammon, J. A. Computer-aided drug-discovery techniques that account for receptor flexibility. *Curr. Opin. Pharmacol.* **2010**, *10* (6), 770–774.

- (22) Colizzi, F.; Perozzo, R.; Scapozza, L.; Recanatini, M.; Cavalli, A. Single-Molecule Pulling Simulations Can Discern Active from Inactive Enzyme Inhibitors. *J. Am. Chem. Soc.* **2010**, *132* (21), 7361–7371.

- (23) Shan, Y.; Kim, E. T.; Eastwood, M. P.; Dror, R. O.; Seeliger, M. A.; Shaw, D. E. How Does a Drug Molecule Find Its Target Binding Site? *J. Am. Chem. Soc.* **2011**, *133* (24), 9181–9183.

- (24) Buch, I.; Giorgino, T.; De Fabritiis, G. Complete reconstruction of an enzyme–inhibitor binding process by molecular dynamics simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, DOI: 10.1073/pnas.1103547108.

- (25) Jorgensen, W. L. Efficient Drug Lead Discovery and Optimization. *Acc. Chem. Res.* **2009**, *42* (6), 724–733.

- (26) Gabdoulline, R. R.; Wade, R. C. Simulation of the diffusional association of barnase and barstar. *Biophys. J.* **1997**, *72* (5), 1917–1929.
- (27) Gabdoulline, R. R.; Wade, R. C. Brownian dynamics simulation of protein-protein diffusional encounter. *Methods* **1998**, *14* (3), 329–41.
- (28) Donald, L. E.; McCammon, J. A. Brownian dynamics with hydrodynamic interactions. *J. Chem. Phys.* **1978**, *69* (4), 1352–1360.
- (29) García de la Torre, J.; Huertas, M. L.; Carrasco, B. Calculation of Hydrodynamic Properties of Globular Proteins from Their Atomic-Level Structure. *Biophys. J.* **2000**, *78* (2), 719–730.
- (30) Dastidar, S. G.; Lane, D. P.; Verma, C. S. Modulation of p53 binding to MDM2: computational studies reveal important roles of Tyr100. *BMC Bioinf.* **2009**, *10* (Suppl 15), S6.
- (31) Im, W.; Beglov, D.; Roux, B. Continuum Solvation Model: computation of electrostatic forces from numerical solutions to the Poisson-Boltzmann equation. *Comput. Phys. Commun.* **1998**, *111* (1–3), 59–75.
- (32) Coalson, R.; Beck, T. L. Numerical Methods for Solving Poisson and Poisson-Boltzmann Type Equations. In *Encyclopedia of Computational Chemistry*; von Rague Schleyer, P., Ed.; John-Wiley: New York, 1998; Vol. 3, pp 2086–2100.
- (33) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98* (18), 10037–10041.
- (34) Dolinsky, T. J.; Czodrowski, P.; Li, H.; Nielsen, J. E.; Jensen, J. H.; Klebe, G.; Baker, N. A. PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.* **2007**, *35* (suppl 2), W522–W525.
- (35) Czodrowski, P.; Dramburg, I.; Sotriffer, C. A.; Klebe, G. Development, validation, and application of adapted PEOE charges to estimate pKa values of functional groups in protein–ligand complexes. *Proteins: Struct., Funct., Bioinf.* **2006**, *65* (2), 424–437.
- (36) Gabdoulline, R. R.; Wade, R. C. Protein-protein association: investigation of factors influencing association rates by Brownian dynamics simulations. *J. Mol. Biol.* **2001**, *306* (5), 1139–1155.
- (37) Gabdoulline, R. R.; Wade, R. C. Effective Charges for Macromolecules in Solvent. *J. Phys. Chem.* **1996**, *100* (9), 3868–3878.
- (38) Elcock, A. H.; Gabdoulline, R. R.; Wade, R. C.; McCammon, J. A. Computer simulation of protein-protein association kinetics: acetylcholinesterase-fasciculin. *J. Mol. Biol.* **1999**, *291* (1), 149–162.
- (39) Hummer, G. Molecular binding: Under water's influence. *Nat. Chem.* **2010**, *2* (11), 906–7.
- (40) Setny, P.; Baron, R.; McCammon, J. A. How Can Hydrophobic Association Be Enthalpy Driven?. *J. Chem. Theory Comput.* **2010**, *6* (9), 2866–2871.
- (41) Seco, J.; Luque, F. J.; Barril, X. Binding site detection and druggability index from first principles. *J. Med. Chem.* **2009**, *52* (8), 2363–71.
- (42) Brooks, C. L.; Onuchic, J. N.; Wales, D. J. Taking a Walk on a Landscape. *Science* **2001**, *293* (5530), 612–613.
- (43) Huang, D.; Caffisch, A., The free energy landscape of small molecule unbinding. *PLoS Comput. Biol.* **2011**, *7* (2), e1002002.
- (44) Conn, P. J.; Christopoulos, A.; Lindsley, C. W. Allosteric modulators of GPCRs: a novel approach for the treatment of CNS disorders. *Nat. Rev. Drug Discovery* **2009**, *8* (1), 41–54.
- (45) Neckers, L.; Ivy, S. P. Heat shock protein 90. *Curr. Opin. Oncol.* **2003**, *15* (6), 419–24.
- (46) Sharp, S. Y.; Prodromou, C.; Boxall, K.; Powers, M. V.; Holmes, J. L.; Box, G.; Matthews, T. P.; Cheung, K.-M. J.; Kalusa, A.; James, K.; Hayes, A.; Hardcastle, A.; Dymock, B.; Brough, P. A.; Barril, X.; Cansfield, J. E.; Wright, L.; Surgenor, A.; Foloppe, N.; Hubbard, R. E.; Aherne, W.; Pearl, L.; Jones, K.; McDonald, E.; Raynaud, F.; Eccles, S.; Drysdale, M.; Workman, P. Inhibition of the heat shock protein 90 molecular chaperone in vitro and in vivo by novel, synthetic, potent resorcinyl pyrazole/isoxazole amide analogues. *Mol. Cancer Ther.* **2007**, *6* (4), 1198–1211.
- (47) Coe, D.; Armer, R.; Ratcliffe, A. Medicines research: current trends and case histories. Highlights of the Society for Medicines Research Symposium. *Drugs Future* **2009**, *34* (3), 247–254.
- (48) Moll, U. M.; Petrenko, O. The MDM2-p53 Interaction. *Mol. Cancer Res.* **2003**, *1* (14), 1001–1008.
- (49) Brown, C. J.; Cheok, C. F.; Verma, C. S.; Lane, D. P. Reactivation of p53: from peptides to small molecules. *Trends Pharmacol. Sci.* **2010**, *32* (1), 53–62.
- (50) Lane, D. P.; Hupp, T. R. Drug discovery and p53. *Drug Discovery Today* **2003**, *8* (8), 347–355.
- (51) Murray, J. K.; Gellman, S. H. Targeting protein-protein interactions: lessons from p53/MDM2. *Biopolymers* **2007**, *88* (5), 657–86.
- (52) Joseph, T. L.; Madhumalar, A.; Brown, C. J.; Lane, D. P.; Verma, C. Differential binding of p53 and nutlin to MDM2 and MDMX: Computational studies. *Cell Cycle* **2010**, *9* (6), 1167–81.
- (53) Walensky, L. D.; Kung, A. L.; Escher, I.; Malia, T. J.; Barbuto, S.; Wright, R. D.; Wagner, G.; Verdine, G. L.; Korsmeyer, S. J. Activation of Apoptosis in Vivo by a Hydrocarbon-Stapled BH3 Helix. *Science* **2004**, *305*, 1466–1470.
- (54) Guo, Z.; Mohanty, U.; Noehre, J.; Sawyer, T. K.; Sherman, W.; Krilov, G. Probing the α -Helical Structural Stability of Stapled p53 Peptides: Molecular Dynamics Simulations and Analysis. *Chem. Biol. Drug Des.* **2010**, *75* (4), 348–359.
- (55) Joseph, T. L.; Lane, D.; Verma, C. S. Stapled peptides in the p53 pathway: Computer simulations reveal novel interactions of the staples with the target protein. *Cell Cycle* **2010**, *9* (22), 4560–4568.
- (56) Okazaki, K.-i.; Takada, S. Dynamic energy landscape view of coupled binding and protein conformational change: Induced-fit versus population-shift mechanisms. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105* (32), 11182–11187.

QM/MM Reweighting Free Energy SCF for Geometry Optimization on Extensive Free Energy Surface of Enzymatic Reaction

Takahiro Kosugi and Shigehiko Hayashi*

Department of Chemistry, Graduate School of Science, Kyoto University, Kyoto 606-8502, Japan

Supporting Information

ABSTRACT: We developed a quantum mechanical/molecular mechanical (QM/MM) free energy geometry optimization method by which the geometry of a quantum chemically treated (QM) molecule is optimized on a free energy surface defined with thermal distribution of the surrounding molecular environment obtained by molecular dynamics simulation with a molecular mechanics (MM) force field. The method called QM/MM reweighting free energy self-consistent field combines a mean field theory of QM/MM free energy geometry optimization developed by Yamamoto (Yamamoto, T. *J. Chem. Phys.* **2008**, *129*, 244104) with a reweighting scheme for updating the MM distribution introduced by Hu et al. (Hu, H., et al. *J. Chem. Phys.* **2008**, *128*, 034105) and features high computational efficiency suitable for exploring the reaction free energy surface of extensive protein conformational space. The computational efficiency with improved treatment of a long-range electrostatic (ES) interaction using the Ewald summation technique permits one to take into account global conformational relaxation of an entire protein of an enzyme in the free energy geometry optimization of its reaction center. We applied the method to an enzymatic reaction of a substrate complex of psychrophilic α -amylase from Antarctic bacterium *Pseudoalteromonas haloplanktis* and succeeded in geometry optimizations of the reactant and the product of the catalytic reaction that involve large conformational changes of protein loops adjacent to the reaction center on time scales reaching sub-microseconds. We found that the adjacent loops in the reactant and the product form in different conformations and produce catalytic ES potentials on the reaction center.

INTRODUCTION

Catalytic reaction mechanisms in enzymes have been interesting and important topics in chemistry and biology.^{1,2} Theoretically, a widely used technique to study enzymatic reactions is a combined quantum mechanical/molecular mechanical (QM/MM) method.^{3–7} In this technique, the enzymatic reaction center is described quantum mechanically, while the surrounding environment of biomacromolecules such as proteins is treated by using a molecular mechanics (MM) force field. The method allows one to take into account complex environmental effects of enzymes for catalysis very efficiently, providing powerful means for studying enzymatic chemical reactions.

Despite the success of the QM/MM approach, however, the method suffers from a difficulty due to the high computational cost of the included QM part. Because of the time-consuming QM calculation, the potential energy profile in only one conformation of a protein is examined in a conventional QM/MM procedure, whereas the chemical reaction proceeds in thermal fluctuation. Thus its energetics are characterized in terms of free energy determined by the thermal average of conformations. In particular, such an approach based on the potential energy surface cannot be applied to enzymatic reaction processes where the chemical steps are correlated with protein conformational changes.

Although, in principle, statistical samples of protein conformations can be collected from a molecular dynamics (MD) simulation with a QM/MM Hamiltonian, the high computational cost of the QM/MM method limits severely the sampling calculation, leading to a poor statistical convergence of the sampling. In particular, a very long trajectory calculation for the sampling is

required for protein systems since the structural relaxation of the protein is known to be very slow.^{8–10} Such a slow relaxation motion of the protein is also suggested to be coupled with a local event of reaction.¹¹ The relaxation time could be tens of picoseconds,¹¹ which is more than 2 orders of magnitude longer than that of water solvent,¹² indicating that a much longer MD calculation for protein than for water solution is necessary for obtaining a properly converged statistical sampling.

It should be noted that use of an adequate QM method in the QM/MM treatment is also crucial for accurate evaluation of the reaction profile. Since catalytic reaction centers are often highly polar, extended QM molecular regions with large basis functions are necessary for a description of the complex electronic nature including polarization and charge transfer. An increase in the computational cost of the QM method for the complex catalytic reaction centers therefore introduces a serious dilemma between the accuracies of the QM description and the statistical sampling.

Several QM/MM methodologies that can take into account thermal fluctuation of a protein conformation for the examination of an enzymatic reaction profile have been developed. MD simulations with QM/MM Hamiltonians have been carried to evaluate reaction free energy surfaces along reaction coordinates. In such straightforward approaches, to improve the convergence of statistical samplings, QM/MM Hamiltonians are approximated with semiempirical methods^{13–18} or empirical potential energy functions^{19–21} of which parameters were determined by more accurate QM or QM/MM methods.

Received: August 20, 2011

Published: November 09, 2011

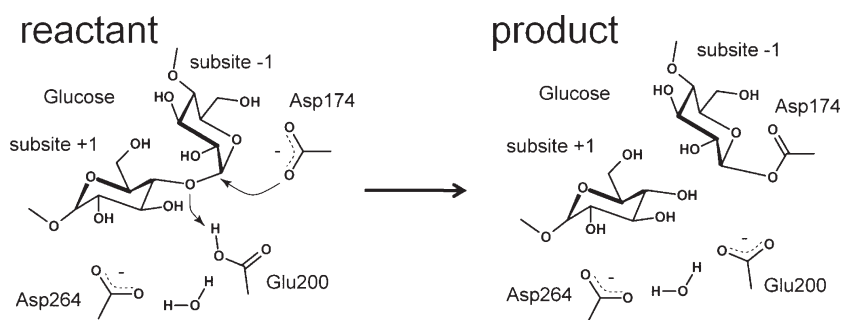


Figure 1. Reaction scheme of cleavage of the α -glycosidic bond in α -amylase catalysis studied in the present study. The product state corresponds to an intermediate of the overall α -amylase catalytic reaction proposed previously.³⁶ Numbers, +1 and -1, indicate the subsites of the substrate.

The other approaches, i.e., QM/MM free energy geometry optimizations,^{22–30} have also been developed extensively. In these approaches, reaction paths are determined by geometry optimizations of enzymatic reaction groups represented by QM/MM Hamiltonians on free energy surfaces defined by statistical samples of the surrounding protein conformations described by MM force fields. Although the approaches omit explicit sampling of the conformations of QM molecules, the omission separates the time-consuming QM calculation and the MM conformational sampling one, enabling one to utilize highly accurate ab initio QM methodologies directly for the description of large QM molecules and to enrich statistical samples of the surrounding protein conformations. However, the convergence of the statistical sampling for enzymes has not been well assessed since the methodologies developed so far are still not efficient enough to obtain sufficient statistical samples for protein systems with very slow conformational relaxation.

In the present study, we developed an efficient QM/MM free energy geometry optimization method that combines a method based on a mean field approximation developed by Yamamoto³⁰ with a reweighting update scheme for the statistical ensemble of a protein conformation introduced by Yang and co-workers.²⁵ Unlike others, the present method, called the QM/MM reweighting free energy self-consistent field (QM/MM-RWFE-SCF), determines a fully variationally electronic wave function of the QM molecules in the surrounding molecular field represented in a mean field manner, leading to a great efficiency in the computational procedure and an improvement of convergence behavior of the geometry optimization necessary for examination of the protein systems. We also incorporated the Ewald summation technique in evaluation of the QM/MM electrostatic (ES) interaction for a periodic boundary condition (PBC) system. In addition to improved accuracy of the QM/MM ES interaction with full electrostatics, the Ewald method provides a consistent description of the ES interaction of the QM/MM method with that of existing highly sophisticated MD program packages, enabling one to obtain efficiently sufficient statistical samples of the protein conformations by the latter.

The QM/MM-RWFE-SCF method was applied to an enzymatic reaction of psychrophilic α -amylase from the Antarctic bacterium *Pseudoalteromonas haloplanktis*. This enzyme and its homologues exhibit a notable temperature dependence for the enzymatic activity, which is suggested to originate from the difference in protein structural flexibility.^{31–35} For the first step to understanding the role of the protein structural flexibility in the enzymatic catalysis, we determined free energetically optimal structures of the active site in a catalytic reaction step schematically

depicted in Figure 1. The reaction, cleavage of the α -glycosidic bond, is the first step of the overall catalytic reaction process proposed previously,³⁶ and thus the state labeled as the product in the present study corresponds to an intermediate state of the overall catalysis.

The high efficiency provided by the present method drastically increases the sizes of the enzyme systems and protein conformational samples treated; we identified the optimized structures of the QM catalytic site described by an ab initio QM method with more than 600 basis functions on free energy surfaces of the conformational samples of the MM protein consisting of more than 68 000 atoms obtained by MD simulations for tens of nanoseconds. In particular, we succeeded in optimizing a structure with the MM conformational samples obtained by a MD trajectory for ~ 90 ns, which involves large and slow conformational changes at a loop adjacent to the catalytic reaction site. The calculations demonstrate critical importance of the sufficient conformational sampling by a long MD simulation.

THEORY

A free energy functional of the QM Born–Oppenheimer (BO) electronic wave function, $\Psi(\mathbf{r};\mathbf{R},\mathbf{X})$, is introduced³⁰ as

$$F[\Psi] = -\beta^{-1} \ln \int d\mathbf{R} d\mathbf{X} \exp(-\beta E[\Psi(\mathbf{r};\mathbf{R},\mathbf{X});\mathbf{R},\mathbf{X}]) \quad (1)$$

where \mathbf{r} , \mathbf{R} , and \mathbf{X} are coordinates of electrons, QM atoms, and MM atoms, respectively, and $E[\Psi(\mathbf{r};\mathbf{R},\mathbf{X});\mathbf{R},\mathbf{X}]$ is the expectation value of the total energy of the QM/MM system:

$$E[\Psi(\mathbf{r};\mathbf{R},\mathbf{X});\mathbf{R},\mathbf{X}] = \langle \Psi(\mathbf{r};\mathbf{R},\mathbf{X}) | \hat{H} | \Psi(\mathbf{r};\mathbf{R},\mathbf{X}) \rangle_{\mathbf{r}} \quad (2)$$

\hat{H} is the total Hamiltonian,

$$\hat{H}(\mathbf{r},\mathbf{R},\mathbf{X}) = \hat{H}^0(\mathbf{r},\mathbf{R}) + \hat{H}^{\text{QM-MM}}(\mathbf{r},\mathbf{R},\mathbf{X}) + E^{\text{MM}}(\mathbf{X}) \quad (3)$$

where $\hat{H}^0(\mathbf{r},\mathbf{R})$ is the gas electronic Hamiltonian of the QM molecules, $E^{\text{MM}}(\mathbf{X})$ is the energy function of the MM ones, and $\hat{H}^{\text{QM-MM}}(\mathbf{r},\mathbf{R},\mathbf{X})$ represents the interaction between the QM and MM parts. The QM–MM interaction is usually given by a sum of the ES nonbonding interaction, $\hat{H}_{\text{ES}}^{\text{QM-MM}}(\mathbf{r},\mathbf{R},\mathbf{X})$, and the interaction other than ES, $E_{\text{non-ES}}^{\text{QM-MM}}(\mathbf{R},\mathbf{X})$, such as the Lenard-Jones nonbonding one and the intramolecular bonding potentials:

$$\hat{H}^{\text{QM-MM}}(\mathbf{r},\mathbf{R},\mathbf{X}) = \hat{H}_{\text{ES}}^{\text{QM-MM}}(\mathbf{r},\mathbf{R},\mathbf{X}) + E_{\text{non-ES}}^{\text{QM-MM}}(\mathbf{R},\mathbf{X}) \quad (4)$$

Note that the QM electrons interact with effective point charges of the MM atoms through the ES interaction, whereas the non-ES interaction is independent of the QM electronic coordinates and is described solely by MM force field parameters. In a standard QM/MM scheme, the ES interaction is given as

$$\hat{H}_{\text{ES}}^{\text{QM-MM}}(\mathbf{r}, \mathbf{X}) = \sum_i^{N_{\text{elec}}} \sum_{\alpha}^{N_{\text{MM}}} \frac{Q_{\alpha}}{|\mathbf{r}_i - \mathbf{X}_{\alpha}|} \quad (5)$$

where N_{elec} and N_{MM} are the numbers of electrons and MM atoms, respectively, and Q_{α} represents the effective charges of the MM atoms.

Taking variation of the free energy functional with respect to the QM wave function, i.e., $\delta F[\Psi]/\delta\Psi = 0$, with a constraint for the normalization of the QM wave function, $\langle\Psi|\Psi\rangle = 1$, provides the variational condition which the optimal QM wave function, $\tilde{\Psi}$, has to satisfy as

$$\int d\mathbf{R} d\mathbf{X} [\hat{H}|\tilde{\Psi}(\mathbf{r}; \mathbf{R}, \mathbf{X})\rangle - E|\tilde{\Psi}(\mathbf{r}; \mathbf{R}, \mathbf{X}); \mathbf{R}, \mathbf{X}\rangle] |\tilde{\Psi}(\mathbf{r}; \mathbf{R}, \mathbf{X})\rangle \times \exp(-\beta E[\tilde{\Psi}(\mathbf{r}; \mathbf{R}, \mathbf{X}); \mathbf{R}, \mathbf{X}]) = 0 \quad (6)$$

The condition indicates that the optimal free energy is given with the BO wave functions and their energies variationally determined at conformations of \mathbf{R} and \mathbf{X} in the Boltzmann distribution of the total energies, $E[\tilde{\Psi}(\mathbf{r}; \mathbf{R}, \mathbf{X}); \mathbf{R}, \mathbf{X}]$. A sampling procedure that fulfills the condition is a straightforward QM/MM MD simulation where the time evolution of the trajectory is computed with the QM/MM wave function and its energy variationally determined at each step of the trajectory calculation. Unfortunately, such a straightforward QM/MM MD simulation, especially based on the ab initio QM method, is computationally impractical for a QM/MM system with a large QM region. As mentioned above, the large QM size is a crucial factor for accurate description of the QM/MM system. Furthermore, the high computational cost limits the sampling time of the trajectory, leading to a poor convergence of the statistical sampling for the thermal distribution.

To reduce the sampling cost, thermal distribution of the QM coordinates, \mathbf{R} , is omitted, and instead, the free energy functional is optimized with respect to \mathbf{R} . This approximate procedure is the so-called free energy geometry optimization.^{22,23} For this purpose, the free energy surface, which is an explicit function of \mathbf{R} , is defined as

$$F[\Psi; \mathbf{R}] = -\beta^{-1} \ln \int d\mathbf{X} \exp(-\beta E[\Psi(\mathbf{r}; \mathbf{R}, \mathbf{X}); \mathbf{R}, \mathbf{X}]) \quad (7)$$

Gradients on the free energy surface are then derived as the averaged energy gradients

$$\frac{\partial F[\tilde{\Psi}; \mathbf{R}]}{\partial \mathbf{R}} = \left\langle \frac{\partial E[\tilde{\Psi}(\mathbf{r}; \mathbf{R}, \mathbf{X}); \mathbf{R}, \mathbf{X}]}{\partial \mathbf{R}} \right\rangle_{\mathbf{X}, E[\tilde{\Psi}(\mathbf{r}; \mathbf{R}, \mathbf{X}); \mathbf{R}, \mathbf{X}]} \quad (8)$$

where $\langle \dots \rangle_{\mathbf{X}, E[\tilde{\Psi}(\mathbf{r}; \mathbf{R}, \mathbf{X}); \mathbf{R}, \mathbf{X}]}$ indicates the thermal average over \mathbf{X} with its thermal distribution obtained for the energy, $E[\tilde{\Psi}(\mathbf{r}; \mathbf{R}, \mathbf{X}); \mathbf{R}, \mathbf{X}]$. The procedure also requires calculations of the optimal QM/MM energy and its forces on \mathbf{R} and \mathbf{X} at every MM conformation so that the computational cost is still very demanding.

The computational cost of the free energy geometry optimization procedure is drastically reduced by introducing the mean field approximation for the electronic wave function where the

explicit dependence on \mathbf{X} is neglected, i.e., $\Psi(\mathbf{r}; \mathbf{R}, \mathbf{X}) \rightarrow \Psi_{\text{MF}}(\mathbf{r}; \mathbf{R})$. The mean field free energy functional is then written as

$$F[\Psi_{\text{MF}}; \mathbf{R}] = E^0[\Psi_{\text{MF}}(\mathbf{r}, \mathbf{R}); \mathbf{R}] - \beta^{-1} \ln \int d\mathbf{X} \exp(-\beta[E^{\text{QM-MM}}[\Psi_{\text{MF}}(\mathbf{r}; \mathbf{R}); \mathbf{R}, \mathbf{X}] + E^{\text{MM}}(\mathbf{X})]) \quad (9)$$

where E^0 and $E^{\text{QM-MM}}$ are expectation values of the gas Hamiltonian, $\hat{H}^0(\mathbf{r}, \mathbf{R})$, and the QM–MM interaction, $\hat{H}^{\text{QM-MM}}(\mathbf{r}, \mathbf{R}, \mathbf{X})$, respectively. A Hartree–Fock equation that determines the optimal mean field wave function and energy can be derived by taking variation of the free energy functional with respect to molecular orbitals. The Fock operator is expressed as

$$\hat{f}^{\text{QM/MM}}(\mathbf{r}; \mathbf{R}) = \hat{f}^0(\mathbf{r}; \mathbf{R}) + \langle \hat{f}_{\text{ES}}^{\text{QM-MM}}(\mathbf{r}; \mathbf{R}, \mathbf{X}) \rangle_{\mathbf{X}, E[\Psi_{\text{MF}}(\mathbf{r}; \mathbf{R}); \mathbf{R}, \mathbf{X}]} \quad (10)$$

where $\hat{f}^0(\mathbf{r}; \mathbf{R})$ is the gas Fock operator and $\hat{f}_{\text{ES}}^{\text{QM-MM}}(\mathbf{r}; \mathbf{R}, \mathbf{X})$ represents the ES interaction of an electron with the MM effective point charges. The Fock term corresponding to the Hamiltonian of eq 5 is expressed as

$$\hat{f}_{\text{ES}}^{\text{QM-MM}}(\mathbf{r}, \mathbf{X}) = \sum_{\alpha}^{N_{\text{MM}}} \frac{Q_{\alpha}}{|\mathbf{r} - \mathbf{X}_{\alpha}|} \quad (11)$$

As eq 10 indicates, the mean field treatment allows one to determine the optimal wave function by solving only one variational problem with the averaged operator, which is much less time-consuming than QM/MM MD simulations where the variational solution has to be obtained at every MM conformation. However, for the standard QM–MM ES interaction operator, the treatment requires calculation of one-electron integrals at every MM conformation, which is still computationally demanding. We therefore introduced the charge operator, $\hat{q}_A(\mathbf{r}, \mathbf{R})$, the expectation value of which gives an effective point charge of a QM atom:

$$q_A(\mathbf{d}, \mathbf{R}) = \langle \Psi_{\text{MF}}(\mathbf{r}, \mathbf{R}) | \hat{q}_A(\mathbf{r}, \mathbf{R}) | \Psi_{\text{MF}}(\mathbf{r}, \mathbf{R}) \rangle_{\mathbf{r}} \quad (12)$$

where \mathbf{d} is the one-electron density matrix. In the present study, the RESP charge operator³⁷ was employed for the charge operator. The QM–MM ES interaction energy is then expressed as

$$E_{\text{ES}}^{\text{QM-MM}}(\mathbf{d}; \mathbf{R}, \mathbf{X}) = \sum_A^{N_{\text{QM}}} q_A(\mathbf{d}; \mathbf{R}) V_A(\mathbf{R}, \mathbf{X}) \quad (13)$$

where $V_A(\mathbf{R}, \mathbf{X})$ is the ES potential of the MM atoms on the QM atom A and N_{QM} is the number of atoms in the QM region. Details on the RESP charge operator and the ES potential are found in ref 37. The averaged Fock operator term of the ES interaction is given as³⁰

$$\langle \hat{f}_{\text{ES}}^{\text{QM-MM}}(\mathbf{r}; \mathbf{R}, \mathbf{X}) \rangle_{\mathbf{X}, E[\Psi_{\text{MF}}(\mathbf{r}; \mathbf{R}); \mathbf{R}, \mathbf{X}]} = \sum_A^{N_{\text{QM}}} \hat{q}_A(\mathbf{r}; \mathbf{R}) \langle V_A(\mathbf{R}, \mathbf{X}) \rangle_{\mathbf{X}, E[q(\mathbf{d}, \mathbf{R}); \mathbf{R}, \mathbf{X}]} \quad (14)$$

One can see in the equation that the ES potentials are independent of the coordinates of electrons so that the ensemble average is taken only with the classical variables \mathbf{R} and \mathbf{X} . Hence, the ensemble average does not include evaluation of the one-electron integrals, which drastically reduces the computational cost.

The ensemble average in eq 14, however, still suffers from a computational difficulty. For the ensemble average over \mathbf{X} , a classical MD simulation of \mathbf{X} is carried with a given $q_A(\mathbf{d}, \mathbf{R})$ and \mathbf{R}

of the QM atoms. Hence, the classical MD simulation has to be performed at every step of SCF and geometry optimization cycles where \mathbf{d} and \mathbf{R} are updated, respectively. Such frequent MD samplings require a very demanding computational cost since tight convergence of the ensemble for each MD sampling needs to be imposed for stable convergence of the SCF and the geometry optimization. In a computational scheme developed by Yamamoto,³⁰ a micro-iteration procedure which avoids the frequent MD samplings is introduced. During the micro-iteration of SCF and geometry optimization, the ensemble of \mathbf{X} is unchanged; even \mathbf{d} and \mathbf{R} are updated. Then a macro-iteration for the update of the ensemble of \mathbf{X} with \mathbf{d} and \mathbf{R} converged at the micro-iteration is carried out in order to achieve the global convergence. However, as seen below, the procedure may slow the global convergence because the update of \mathbf{d} and \mathbf{R} does not proceed on the proper free energy surface defined by the functional of eq 9 at most steps of the micro-iteration.

In the QM/MM-RWFE-SCF method developed in the present study, we have incorporated a reweighting update of the ensemble of \mathbf{X} employed in QM/MM-MFEP methods by Yang and co-workers²⁵ into the MF free energy functional formalism by Yamamoto.³⁰ By introducing a reference QM system with an electronic wave function, $\Psi_{\text{MF,ref}}(\mathbf{r}_{\text{ref}}; \mathbf{R}_{\text{ref}})$, and a geometry, \mathbf{R}_{ref} , the free energy difference from the reference value, $\Delta F[\Psi_{\text{MF}}; \mathbf{R}] = F[\Psi_{\text{MF}}; \mathbf{R}] - F[\Psi_{\text{MF,ref}}; \mathbf{R}_{\text{ref}}]$, is rewritten as

$$\begin{aligned} \Delta F[\Psi_{\text{MF}}; \Psi_{\text{MF,ref}}; \mathbf{R}; \mathbf{R}_{\text{ref}}] &= E^0[\Psi_{\text{MF}}(\mathbf{r}; \mathbf{R}); \mathbf{R}] - E^0[\Psi_{\text{MF,ref}}(\mathbf{r}_{\text{ref}}; \mathbf{R}_{\text{ref}}); \mathbf{R}_{\text{ref}}] \\ &\quad - \beta^{-1} \ln \langle \exp(-\beta \Delta E^{\text{QM-MM}}(\mathbf{d}, \mathbf{d}_{\text{ref}}; \mathbf{R}, \mathbf{R}_{\text{ref}}; \mathbf{X})) \rangle_{\mathbf{X}, E[\mathbf{q}(\mathbf{d}_{\text{ref}}, \mathbf{R}_{\text{ref}}); \mathbf{R}_{\text{ref}}, \mathbf{X}]} \end{aligned} \quad (15)$$

where \mathbf{d}_{ref} is the one electron density of the reference electronic wave function and $\Delta E^{\text{QM-MM}}(\mathbf{d}, \mathbf{d}_{\text{ref}}; \mathbf{R}, \mathbf{R}_{\text{ref}}; \mathbf{X})$ is the QM–MM interaction energy difference from the reference value. The reference density and coordinates are better chosen to be similar to the optimal ones in order to attain smaller QM–MM interaction energy differences which provide quicker convergence of the statistical sampling, although they can be arbitrary in principle. In the present study, we adopted the ones obtained by conventional static QM/MM calculations for the initial values. As seen below, the reference density and coordinates are renewed in a sequential sampling. The ensemble average over \mathbf{X} is taken with MD samples obtained for the reference QM system, and thus the update of the ensemble of \mathbf{X} at each step of SCF and geometry optimization cycles is not necessary.

The ensemble average of the ES potential in the Fock operator of eq 14 is also rewritten with \mathbf{d}_{ref} and \mathbf{R}_{ref} as

$$\begin{aligned} \langle V_A(\mathbf{R}, \mathbf{X}) \rangle_{\mathbf{X}, E[\mathbf{q}(\mathbf{d}, \mathbf{R}); \mathbf{R}, \mathbf{X}]} &= \int d\mathbf{X} V_A(\mathbf{R}, \mathbf{X}) \exp[-\beta(\Delta E^{\text{QM-MM}}(\mathbf{d}, \mathbf{d}_{\text{ref}}; \mathbf{R}, \mathbf{R}_{\text{ref}}; \mathbf{X}))] \\ &\quad \times \exp[-\beta(E^{\text{QM-MM}}(\mathbf{d}_{\text{ref}}; \mathbf{R}_{\text{ref}}; \mathbf{X}) + E^{\text{MM}}(\mathbf{X}))] \\ &\quad \Big/ \int d\mathbf{X} \exp[-\beta(\Delta E^{\text{QM-MM}}(\mathbf{d}, \mathbf{d}_{\text{ref}}; \mathbf{R}, \mathbf{R}_{\text{ref}}; \mathbf{X}))] \\ &\quad \times \exp[-\beta(E^{\text{QM-MM}}(\mathbf{d}_{\text{ref}}; \mathbf{R}_{\text{ref}}; \mathbf{X}) + E^{\text{MM}}(\mathbf{X}))] \\ &= \langle V_A(\mathbf{R}, \mathbf{X}) \omega(\mathbf{d}, \mathbf{d}_{\text{ref}}; \mathbf{R}, \mathbf{R}_{\text{ref}}; \mathbf{X}) \rangle_{\mathbf{X}, E[\mathbf{q}(\mathbf{d}_{\text{ref}}, \mathbf{R}_{\text{ref}}); \mathbf{R}_{\text{ref}}, \mathbf{X}]} \end{aligned} \quad (16)$$

where $\omega(\mathbf{d}, \mathbf{d}_{\text{ref}}; \mathbf{R}, \mathbf{R}_{\text{ref}})$ is the reweighting factor:

$$\begin{aligned} \omega(\mathbf{d}, \mathbf{d}_{\text{ref}}; \mathbf{R}, \mathbf{R}_{\text{ref}}; \mathbf{X}) &= \frac{\exp(-\beta \Delta E^{\text{QM-MM}}(\mathbf{d}, \mathbf{d}_{\text{ref}}; \mathbf{R}, \mathbf{R}_{\text{ref}}; \mathbf{X}))}{\langle \exp(-\beta \Delta E^{\text{QM-MM}}(\mathbf{d}, \mathbf{d}_{\text{ref}}; \mathbf{R}, \mathbf{R}_{\text{ref}}; \mathbf{X})) \rangle_{\mathbf{X}, E[\mathbf{q}(\mathbf{d}_{\text{ref}}, \mathbf{R}_{\text{ref}}); \mathbf{R}_{\text{ref}}, \mathbf{X}]} \end{aligned} \quad (17)$$

Because of the reweighting factor, the Fock operator of the ES interaction between the QM and MM regions depends on the electron density, unlike conventional QM/MM methods, so that the matrix element is calculated at each step of the SCF cycle. Extension of the present scheme to other variational methods such as density functional theory (DFT) and complete active space SCF is straightforward.

The reweighting treatment is also applied straightforwardly to the calculation of the mean force given by eq 8. Expression of the derivative of the RESP charge operator given by eq 12 with respect to \mathbf{R} is found in ref 37. It is noteworthy that one does not need to carry out coupled perturbed calculations to evaluate explicitly derivatives of linear-combination-of-atomic-orbitals (LCAO) coefficients of the molecular orbitals with respect to \mathbf{R} . The molecular orbitals are determined variationally for the free energy functional, eq 15, by solving the consistent Fock or Kohn–Sham (KS) equation with the operator, eq 10, with eqs 14 and 16. Hence, the free energy gradient terms including derivatives of the LCAO coefficients can be replaced by terms with the orbital energies and derivatives of the overlap integrals as in conventional energy gradient techniques. The reweighting scheme also allows one to calculate efficiently the Hessian matrix on the free energy surface through the finite differential method, since the same MM conformational samples can be used for the free energy gradient calculations of the slightly displaced QM geometries. The efficient scheme for the calculation of the Hessian matrix enables one to determine a transition state structure and to evaluate zero point energy and vibrational entropy. Those calculations are now ongoing and will be reported elsewhere.

As described above, the reweighting method permits one to update the ensemble of \mathbf{X} without reevaluating it with an MD trajectory calculation. However, the reweighting update of the ensemble of \mathbf{X} is valid only if the ensemble of \mathbf{X} covers a large configurational space of \mathbf{X} . Thus, for limited samples of \mathbf{X} obtained by an MD simulation, the changes of \mathbf{d} and \mathbf{R} from their references, \mathbf{d}_{ref} and \mathbf{R}_{ref} , need to stay small in SCF and geometry optimization; if the deviation of \mathbf{d} and \mathbf{R} from their references is large, only a few samples come to possess dominantly large reweighting factors, eq 17, because of the exponential nature of weighting, leading to a poor ensemble average over \mathbf{X} . It is therefore necessary to perform the sequential sampling²⁵ where an MD trajectory calculation is iteratively carried out at each end of the geometry optimization in order to update the ensemble of \mathbf{X} for renewed \mathbf{d}_{ref} and \mathbf{R}_{ref} until the deviations of \mathbf{d} and \mathbf{R} from the references stay small in the following geometry optimization and consequently a sufficiently large number of the samples come to contribute to the average over \mathbf{X} . Since the behavior of the averaging with the reweighting factors depends strongly on the sample size, careful assessment of the convergence of the sampling is necessary.

In order to evaluate the ES interaction between the QM and MM regions accurately and efficiently, the Ewald method was

implemented in the QM/MM-RWFE-SCF calculation. The ES interaction energy is expressed as

$$V = V^{\text{real}} + V^{\text{rec}} + V^{\text{corr}} \quad (18)$$

where V^{real} and V^{rec} are interaction energies in real and reciprocal spaces, respectively

$$V^{\text{real}} = \sum_n \sum_i^{N_{\text{QM}}} \sum_j^{N_{\text{MM}}} \hat{q}_i q_j \frac{\text{erfc}(\alpha |\mathbf{R}_i - \mathbf{X}_j + \mathbf{L}n|)}{|\mathbf{R}_i - \mathbf{X}_j + \mathbf{L}n|} \quad (19)$$

$$V^{\text{rec}} = \frac{4\pi}{V} \sum_{\mathbf{G} \neq 0} \frac{\exp(-|\mathbf{G}|^2/4\alpha^2)}{|\mathbf{G}|^2} \left\{ \sum_i^{N_{\text{QM}}} \hat{q}_i \cos(\mathbf{G} \cdot \mathbf{R}_i) W_{\cos}(\mathbf{G}) + \sum_i^{N_{\text{QM}}} \hat{q}_i \sin(\mathbf{G} \cdot \mathbf{R}_i) W_{\sin}(\mathbf{G}) \right\} \quad (20)$$

$$W_{\cos}(\mathbf{G}) = \sum_j^{N_{\text{MM}}} q_j \cos(\mathbf{G} \cdot \mathbf{X}_j) \quad (21)$$

$$W_{\sin}(\mathbf{G}) = \sum_j^{N_{\text{MM}}} q_j \sin(\mathbf{G} \cdot \mathbf{X}_j)$$

and V^{corr} is a correction which subtracts self-interaction within the QM–MM boundary region from the term in reciprocal space, eq 20

$$V^{\text{corr}} = - \sum_{(i,j)}^{N_{\text{QM-MM}}} \hat{q}_i q_j \frac{\text{erf}(\alpha |\mathbf{R}_i - \mathbf{X}_j|)}{|\mathbf{R}_i - \mathbf{X}_j|} \quad (22)$$

N_{QM} and N_{MM} are the numbers of atoms in the QM and MM regions, respectively, and $N_{\text{QM-MM}}$ is that of atom pairs included in bonding interactions at the boundary between the QM and MM regions. \hat{q} and q represent the RESP charge operator of the QM atoms and the atomic charge of the MM force field, respectively. α is the screening parameter. \mathbf{L} indicates the box length vector, and V is the box volume. \mathbf{G} is the reciprocal space vector

$$\mathbf{G} = 2\pi \begin{pmatrix} k_x/L_x \\ k_y/L_y \\ k_z/L_z \end{pmatrix} \quad (23)$$

where k_x , k_y , and k_z are integer numbers. Since the RESP charge operator is used for the QM–MM ES interaction, computation of the Ewald interaction terms is straightforward.

For computational simplicity, interaction between the QM regions in different image cells is neglected. Because distance between the nearest neighbor QM regions is large and alignment of the periodic QM images is homogeneous, the QM images add only a constant ES field with negligibly small deviation, as shown in Figures S2 and S3 in the Supporting Information. As the summations for the MM part, $W_{\cos}(\mathbf{G})$ and $W_{\sin}(\mathbf{G})$, do not include the QM atom, these terms are calculated only once at the first step of free energy optimization upon update of the MM ensemble. We used a small screening parameter, $\alpha = 0.13149 \text{ \AA}^{-1}$, and a large cutoff distance of 20 Å for the QM–MM interaction in real space, eq 19, compared with those employed in MD simulations. These parameters give a quick convergence of the time-consuming reciprocal space summation with respect to \mathbf{G} in

eq 20 and thus lead to drastic reduction of the overall computational cost, despite a moderate increase of cost of the QM–MM interaction in real space, eq 19, which is only proportional to the number of atoms in the cutoff range. According to the convergence behavior shown in Figure S4 in the Supporting Information, k_x , k_y , and k_z are set to be from -10 to $+10$, respectively. The present parameters of the Ewald summation were confirmed to give essentially the same description of the ES interaction as that by a particle mesh Ewald (PME) method³⁸ with default parameters of Amber9 program packages;³⁹ the discrepancy of the ES interaction energies for a conformation was evaluated to be within $1.0 \times 10^{-3} \%$.

It is noted that most of recent sophisticated MD programs which execute trajectory calculations very efficiently utilize Ewald-based techniques such as the PME method. The equivalent description of the ES interaction in the QM/MM-RWFE-SCF calculation to that in the MD programs permits one to employ externally the efficient existing MD programs for the conformational sampling of \mathbf{X} in the QM/MM-RWFE-SCF procedure.

The protocol of free energy optimization with the QM/MM RWFE-SCF method is summarized as follows:

- (1) Initial reference coordinates, \mathbf{R}_{ref} and a density matrix, \mathbf{d}_{ref} , of the QM region are given. In this study, results of conventional QM/MM calculation for a cluster system were used as the initial values.
- (2) To obtain an ensemble of \mathbf{X} , a MD simulation is performed with the fixed \mathbf{R}_{ref} and \mathbf{q}_{ref} , which is the RESP charges derived from \mathbf{d}_{ref} .
- (3) The optimal wave function is calculated by the SCF procedure with eq 10. $W_{\cos}(\mathbf{G})$ and $W_{\sin}(\mathbf{G})$ in eq 20 are calculated at the first step of SCF iteration when the ensemble of \mathbf{X} is updated in the sequential sampling procedure and stored throughout a geometry optimization cycle. At each SCF iteration step, the ensemble of \mathbf{X} is updated by eq 16 as the density matrix, i.e., the QM charges, changes and accordingly the ensemble average of the ES potential in the Fock or KS operator of eq 14 is reevaluated.
- (4) Using the optimized wave function, the averaged force is calculated with eq 8 and the QM geometry is updated in a standard geometry optimization manner. The reweighted distribution is used for the calculation of the averaged forces of the QM/MM interaction. If the force is smaller than a convergence criterion, the optimization procedure goes to step 5. Otherwise, the optimization procedure returns to step 3 after the QM geometry is updated.
- (5) If the number of the effective MM samples for the reweighting average is not sufficient, a MD trajectory calculation for the sequential sampling is carried out to update the ensemble of \mathbf{X} as discussed above. The QM reference coordinates and densities, \mathbf{R}_{ref} and \mathbf{d}_{ref} are updated with those optimized at the preceding cycle of steps 3 and 4, and then the procedure returns to step 2 to obtain a new ensemble of \mathbf{X} . If displacements from the reference values upon the following two successive sequential samplings are small and the reweighting averages are taken with sufficient numbers of the MM samples, the optimization procedure is finished.

As seen above, the QM/MM-RWFE-SCF geometry optimization determines the stationary structure on the free energy surface of a given MM ensemble obtained at the end of the

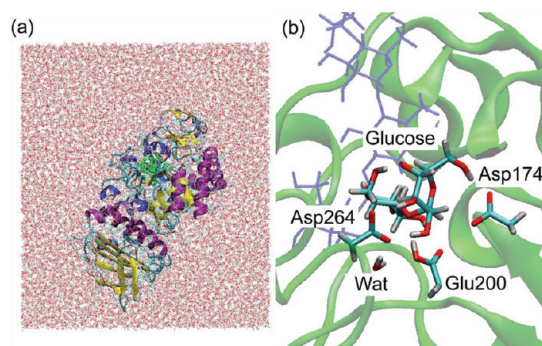


Figure 2. (a) Total simulation system of the QM/MM-RWFE-SCF calculation. The QM region is drawn in green licorice representation. (b) A close-up view of the QM region of the QM/MM-RWFE-SCF calculation. The QM region is drawn in licorice representation.

sequential sampling. For the examination of the reaction free energy profile, the free energy minimum structures of the states in the reaction process such as the reactant and product are therefore determined with different MM ensembles. Hence, the QM/MM-RWFE-SCF calculation cannot solely evaluate the free energy differences between the states. Nevertheless, one can calculate the free energy differences by standard techniques such as the free energy perturbation using the coordinates and the charges of the QM molecule determined by the QM/MM-RWFE-SCF geometry optimizations. The calculation of the reaction free energy profile is now ongoing and will be reported elsewhere.

COMPUTATIONAL DETAIL

The initial protein structure of α -amylase was taken from PDB 1G94.⁴⁰ The structure is of the protein with a saccharide substrate analog compound which occupies subsites -4 to $+3$. We replaced the substrate analog with an amylose, which consists of six α -glucoses in subsites -4 to $+2$. We determined the charge state of acidic residues located at the active site based on a proposed reaction mechanism of α -amylases,³⁶ while standard charge states were assumed for other acidic and basic residues. Hydrogen atoms were added by the LEaP module of AMBER9.³⁹ AMBER ff99 and GLYCAM04 parameter sets⁴¹ were utilized for the protein and glucoses, respectively. For water molecules, the TIP3P model⁴² was used. The protein system was immersed in a water box of $80 \text{ \AA} \times 100 \text{ \AA} \times 90 \text{ \AA}$ (Figure 2) in PBC. To neutralize the system, 16 sodium ions were put in the box. The AMBER9 software suite³⁹ was used for all MD simulations. Long-range ES interactions were treated with the PME method. Short range nonbonded interactions were cut off at 10 \AA .

The QM region employed in the QM/MM calculations is depicted in Figure 2. It consists of two glucoses at the subsites -1 and $+1$, the side chains of two aspartic acids, Asp174 and Asp264, and the side chain from the C_γ atom to the end of Glu200. We used the DFT method with the B3LYP functional for the QM region. The 6-31G* basis set was employed, except for the carboxyl groups of Asp and Glu, where the 6-31+G* basis set was employed. A link atom approach was used to terminate properly chemical bonds at the QM–MM boundaries.³⁷ The convergence criterion of the geometry optimization was set to be 1.0×10^{-3} Hartree/Bohr for the largest component of the gradient for the QM region. The QM/MM-RWFE-SCF method was implemented in the GAMESS program package.⁴³

In order to obtain the initial reference coordinates and electronic density, \mathbf{R}_{ref} and \mathbf{d}_{ref} , conventional QM/MM calculations based on the potential energy surface³⁷ for a sphere cluster system were first carried out. The initial system of the QM/MM calculations was prepared as follows. A 500-ps equilibrium MD simulation with the reactant substrate in a constant-NPT (300 K, 1 atm) ensemble was carried out. The sphere cluster that contains all atoms at less than 40 \AA from the C atom of the scissile bond in the reactant state was taken from the last structure of the MD simulation. Then, an equilibrium 400-ps MD simulation at 300 K was carried out for the sphere cluster system. The residues in the region more than 32 \AA away from the C atom of the scissile bond were fixed in order to keep the shape of the sphere cluster. In the last 200-ps, the temperature was lowered gradually. Finally, an energy minimization was performed until the RMS energy gradient was below 1.0×10^{-4} kcal/mol/ \AA (8.4×10^{-8} Hartree/Bohr). The last structure was used as the initial reactant structure for the QM/MM potential energy geometry optimization. The convergence criterion of the geometry optimization for the MM region was set to be 1.0×10^{-4} Hartree/Bohr for the RMS gradient. We also determined the reference structures and charges of the product state through the QM/MM potential energy geometry optimization from an initial structure obtained by cleaving the α -glycosidic bond of the reactant QM molecule. For the sphere cluster system, no cutoff of the nonbonded interaction was applied throughout the MD and QM/MM calculations.

MD simulations for obtaining the MM ensemble in the sequential sampling of the QM/MM-RWFE-SCF geometry optimization were performed for the PBC system with the fixed reference coordinates and charges of the QM region. A 3-ns trajectory was calculated at each iteration step of the sequential sampling, and the first 1-ns and the last 2-ns trajectories were employed for equilibration and sampling of the MM ensemble, respectively. The MM conformational samples were taken at every 100 fs, and thus the MM ensemble at each step of the sequential sampling was comprised of 20 000 configurations. For the first step of the iteration of the sequential sampling, the MD simulation was carried out from the initial structure where the sphere cluster optimized by the QM/MM calculation described above was inserted into the PBC system (a water box of $80 \text{ \AA} \times 100 \text{ \AA} \times 90 \text{ \AA}$). The total number of atoms in this box is 68 533. The first MD simulation was performed by the SANDAR module under constant-NPT (1 atm, 283 K) conditions for the QM region fixed with the SHAKE algorithm. For MD simulations in the iteration of the sequential sampling after the first step, the trajectories were calculated by the PMEMD module under constant-NVT (283 K) conditions for the space fixed QM region, which drastically accelerates the computation of the trajectory.

RESULTS AND DISCUSSION

Long-Range ES Interaction of the QM Catalytic Site with the Protein. It is well-known that long-range ES interaction is important for protein systems where charge distributions are very inhomogeneous.⁴⁴ Nevertheless, because of computational difficulty, sphere-shaped cluster QM/MM systems were often employed, and long-range ES interactions beyond the cluster sizes were neglected. In order to access the importance of the long-range ES interaction, we compared the ES potentials on the QM sites from the MM surroundings of the cluster system with that of the PBC system with the Ewald method (PBC-Ewald). Figure 3 shows differences in the ES potentials between the cluster systems

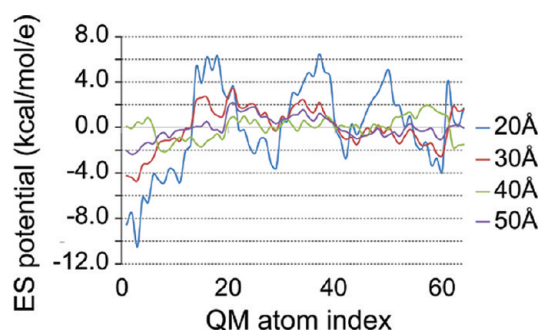


Figure 3. Comparison of ES potentials of the MM region acting on the QM atoms obtained by the Ewald method and cutoff ones. Deviations of ES potentials in the reactant calculated by the residue-based cutoff methods with cutoff distances of 20, 30, 40, and 50 Å from those by the Ewald method are shown. The cutoff distance is defined as the distance from C(40), i.e., the C atom of the scissile bond. The QM atom index is given in Figure S1 in the Supporting Information.

with different cutoff distances and the PBC-Ewald system. The ES potentials for the smaller cluster systems deviate largely from that for the PBC-Ewald system. As the sphere cluster size increases, the ES potentials of the cluster systems converge gradually to those of the PBC-Ewald one, and finally the difference between those of the cluster systems with a cutoff of 50 Å and of the PBC-Ewald system falls within ± 2.0 kcal/mol. This indicates clearly that the long-range ES interaction between the QM and MM regions is important, and the Ewald method provides an accurate description of the interaction.

It should be noted that, although a large cluster system with a cutoff of 50 Å could give an accuracy of the ES interaction comparable with the PBC-Ewald system, the MD sampling simulation in the QM/MM-RWFE-SCF procedure for the large cluster system is much more computationally demanding than that for the PBC-Ewald system with the PME method; the increase of its computational cost for the former is proportional to N^2 (N is the number of atoms treated), whereas that for the latter is proportional to $N \log N$. On an Intel Xeon E5450 computer cluster with 32 cores, the PMEMD module with the PME method for the PBC system executes already much faster than the SANDER module for the sphere cluster system of 30 Å. Furthermore, the cluster system requires a fixed or restrained sphere shell boundary to keep its shape and thus arrests global flexibility of the protein, which may play a role in the catalytic reaction. The PBC-Ewald system is therefore advantageous over the cluster system in terms of both computational accuracy and efficiency.

Convergence Behavior of MM Conformational Distribution Sampled by MD Simulation. As described above, the electronic wave function and the geometry of the QM region are determined under mean fields of the MM conformational distributions sampled by MD simulations. Thus, statistical convergence of the MM distribution is a crucial factor for the stable calculation. We therefore examined carefully the convergence behaviors of the MM distributions. In the QM/MM-RWFE-SCF geometry optimization, convergences of the MM distributions at two different iterations are necessary, i.e., convergence of the geometry optimization in each iteration of the sequential sampling and that of the macro-iteration of the sequential sampling.

First, we assessed the former convergence. In each step of the sequential sampling iteration, the geometry is optimized with a

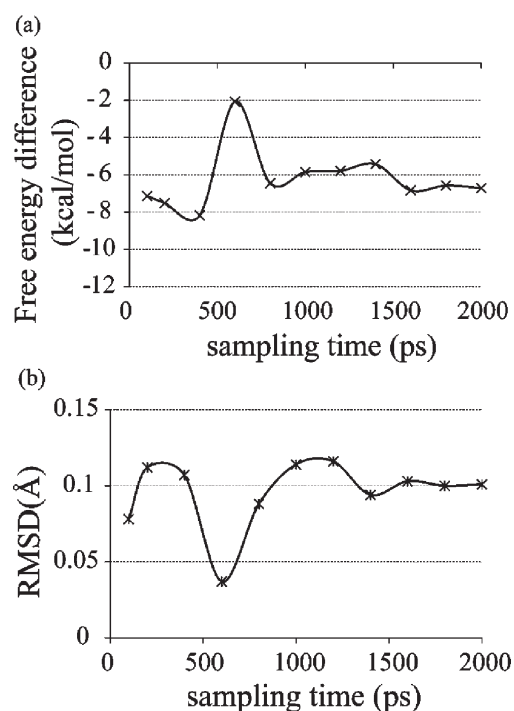


Figure 4. Convergence behaviors of the QM/MM free energy optimization with a single MM conformational distribution in the first step of the sequential sampling iteration along its cumulative sampling time by a MD simulation. Changes of free energy difference given by eq 15 (a) and RMSDs from the initial structure (b) are plotted.

single set of the MM samples. Figure 4 indicates convergences of free energy and root-mean-square deviation (RMSD) of the QM geometry optimized at the first step of the sequential sampling iteration with respect to cumulative MD sampling time for the MM distribution. For the present system, both the free energy and the QM geometry do not converge before 1.5 ns, which is much longer than simulation time of other previous studies of free energy geometry optimization,^{23,25,29,30} i.e., several hundreds of picoseconds or less. It was found that the optimized QM geometries with MM distributions of different cumulative MD sampling times before 1.5 ns differ significantly from one another, which leads to an unstable search for the geometry optimization, whereas after 1.5 ns the geometry optimizations with different MM distributions converge to a single QM geometry. We therefore employed MD trajectories for 2.0 ns for each step of the sequential sampling iteration in this study.

Second, convergence of the overall iteration of the sequential sampling was examined. As described above, the reweighting scheme with limited MM samples suffers from the difficulty of poor averaging when the deviations of \mathbf{d} and \mathbf{R} from their reference values become large. The difficulty can be seen in Figure S5, which shows a histogram of the energy difference, $\Delta E^{\text{QM-MM}}(\mathbf{d}, \mathbf{d}_{\text{ref}}; \mathbf{R}, \mathbf{R}_{\text{ref}}; \mathbf{X})$, in eqs 15 and 17 for the MM samples at the optimized geometry in the first step of the sequential sampling iteration and its average over the reweighted MM distribution

$$\begin{aligned} & \langle \Delta E^{\text{QM-MM}} \rangle_{\mathbf{X}, E[\mathbf{q}(\mathbf{d}, \mathbf{R}); \mathbf{R}, \mathbf{X}]} \\ &= \langle \Delta E^{\text{QM-MM}}(\mathbf{d}, \mathbf{d}_{\text{ref}}; \mathbf{R}, \mathbf{R}_{\text{ref}}; \mathbf{X}) \omega(\mathbf{d}, \mathbf{d}_{\text{ref}}; \mathbf{R}, \mathbf{R}_{\text{ref}}; \mathbf{X}) \rangle_{\mathbf{X}, E[\mathbf{q}(\mathbf{d}_{\text{ref}}, \mathbf{R}_{\text{ref}}); \mathbf{R}_{\text{ref}}, \mathbf{X}]} \end{aligned} \quad (24)$$

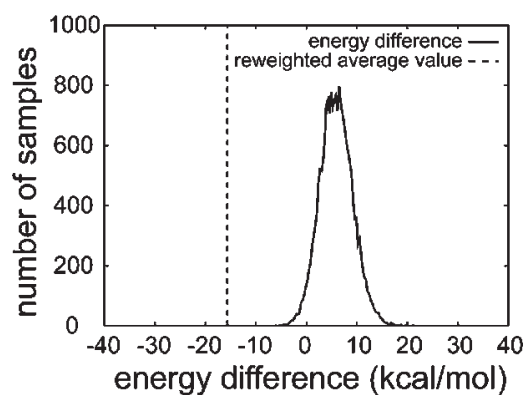


Figure 5. Histogram of the QM–MM interaction energy differences of MM conformational samples, $\Delta E^{\text{QM-MM}}(\mathbf{d}, \mathbf{d}_{\text{ref}}, \mathbf{R}, \mathbf{R}_{\text{ref}}, \mathbf{X})$, and its reweighted average, eq 24, at the end of the geometry optimization cycle in the first step of the sequential sampling.

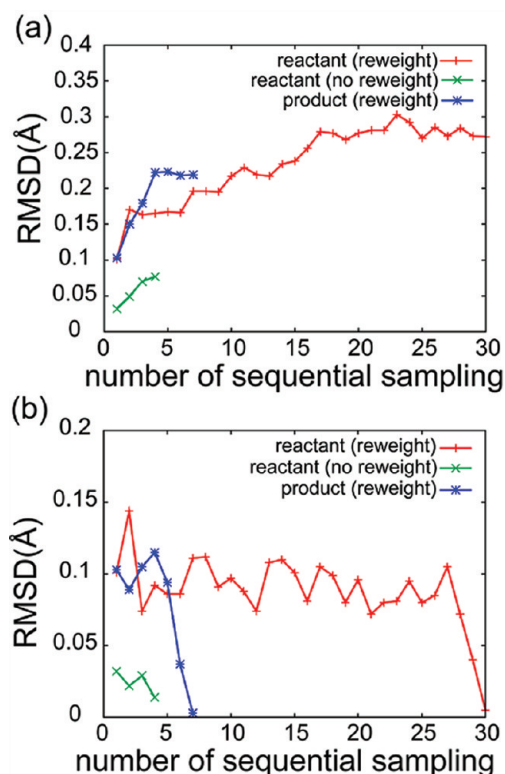


Figure 6. Changes of the QM geometries during the sequential samplings in the reactant and the product. RMSDs from the initial QM coordinates (a) and RMSDs from the QM coordinates optimized at the preceding steps of the sequential samplings (b) are plotted. Red and blue lines indicate RMSDs in the reactant and the product, respectively. Green lines indicate RMSDs of free energy geometry optimizations without the reweighting of the MM distribution in the reactant.

The reweighted average value is far out of the distribution as one MM conformation which exhibits a large energy difference gives a dominantly large reweighting factor. Although the MM conformation is found to provide a strongly stabilizing QM–MM interaction, the statistical averaging over the reweighted distribution represented only by the single MM conformation is no longer valid. This ill behavior of the reweighted average is due to

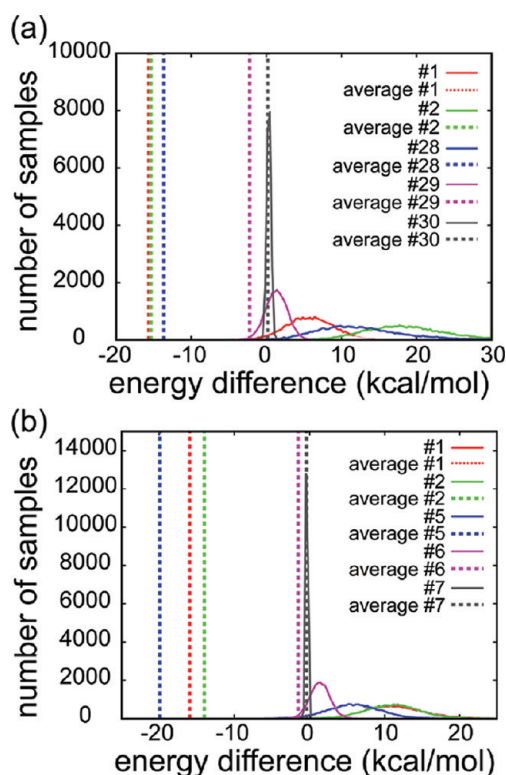


Figure 7. Changes of histogram of the QM–MM interaction energy differences of MM conformational samples, $\Delta E^{\text{QM-MM}}(\mathbf{d}, \mathbf{d}_{\text{ref}}, \mathbf{R}, \mathbf{R}_{\text{ref}}, \mathbf{X})$, and its reweighted average, eq 24, during the sequential samplings in the reactant (a) and the product (b). The histogram and the average are calculated with QM coordinates and charges at the end of the geometry optimization in each step of the sequential samplings. Histograms and reweighted averages of the first two and the last three steps of the sequential samplings are shown.

limited MM samples which fail to cover the MM distributions for the updated \mathbf{d} and \mathbf{R} during the geometry optimization. Naturally, the fewer MM samples obtained by the shorter MD trajectories tested above were observed to lead to ill behavior as well. The sequential sampling which redistributes the MM samples therefore needs to be continued until the ill behavior of the reweighted average disappears. The high computational efficiency featured in the present method is therefore a prerequisite for obtaining sufficient MM samples that avoid the ill behavior of the reweighting scheme.

Figure 6 depicts RMSD changes of the QM coordinates along the sequential sampling iterations. The QM geometries undergo changes from the initial ones as the sequential samplings proceed. Then, the geometry optimizations converged at 30 and 7 steps of the sequential samplings for the reactant and product states, respectively. Since the MD trajectory at each step of the sequential sampling is calculated for 3 ns (1 ns for equilibration and 2 ns for the sampling of the MM distribution), the MD simulations were carried out for 90 and 21 ns in total for the reactant and the product. The reason for the very long MD simulation time required for the convergence in the reactant state is discussed later. As seen in RMSDs from the previous optimized QM coordinates (Figure 6b), the movements of the QM coordinates become small in the last two steps of the sequential samplings, indicating clear convergence behavior of the MM distribution. Figure 7 shows the histograms of the energy difference and their

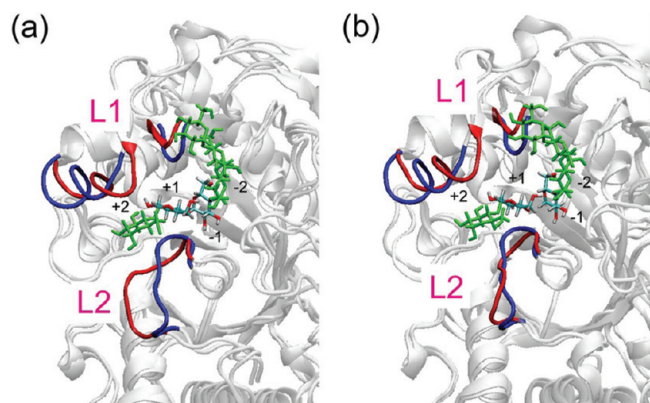


Figure 8. Conformational changes of the protein loops, L1 and L2, adjacent to the substrate binding site observed in the free energy geometry optimizations. The loops undergo large conformational transitions from the initial MM structure obtained by the QM/MM potential energy geometry optimization (blue) to the final structure of the free energy geometry optimization (red) in the reactant (a) and the product (b). Reweighted average structures are shown for the free energetically optimized ones. The L1 and L2 loops correspond to regions 5 and 9, respectively, defined in a previous paper.³⁵ The substrates are depicted in licorice representation, and their QM and MM regions are drawn in colors based on the atom type and in green, respectively. Numbers, -2 to $+2$, are the subsite indices of the substrate.

averaged values at the early and final steps of the sequential sampling iterations. For both of the reactant and product states, the reweighted averages exhibit the ill behavior seen above before the last two steps of the sequential samplings. On the other hand, in the last two steps, the reweighted averages stay in well-distributed regions of the histograms. The averaging with the reweighted distribution therefore becomes statistically valid with the MM distribution at the end of the sequential sampling.

Finally, the validity of the mean field approximation for the electronic wave function is assessed with the MM distribution of the last sampling. We compared the QM charges and the QM/MM energy (the QM energy plus the QM–MM interaction energy) obtained in the mean ES field to those evaluated without the mean field approximation. The latter are obtained through a reweighted average of the quantities which are determined by a series of QM/MM calculations for individual MM configurations in the distribution. For comparison, we limited the number of MM configuration for the ensemble averages to 2000 out of 20 000 of the MM distribution since the series of QM/MM calculations for the average evaluation without the mean field approximation is very time-consuming. We recalculated the mean field quantities using a QM/MM-RWFE-SCF calculation with the small MM ensemble and confirmed that the reweighted average of the QM–MM interaction energy differences stays in a well-distributed region of their histogram for the small ensemble.

Figure S5 (Supporting Information) compares the QM charges determined with and without the mean field approximation. The mean field charges are in accord with those without the mean field approximation. The maximum value and the standard deviation of the error are 7.07×10^{-4} and 2.19×10^{-4} , respectively. The error of the QM/MM energy, 0.48 kcal/mol, is also reasonably small. The mean field approximation therefore provides a reasonable description for the complex protein system as well as for simple solutions systems reported previously.^{30,45,46}

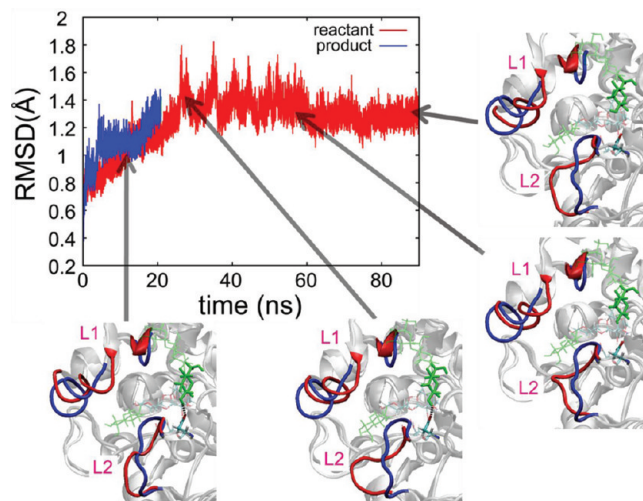


Figure 9. Time evolution of the protein conformational changes during the sequential samplings. RMSDs of the MM structures from the initial ones in the reactant and the product are shown. Representative snapshots of the structure during the sequential sampling in the reactant are also depicted. L1 and L2 protein loops drawn in blue and red are the initial structure and the snapshot structures, respectively. The substrate and Asp264 are drawn in licorice representation, and their QM and MM regions are depicted in colors based on the atom type and in green, respectively. A dashed line indicates a hydrogen bond between glucose (-2) and Asp264 drawn in thick licorice representation, which undergoes dissociation around 60 ns.

Large Conformational Changes of Protein Found in the QM/MM-RWFE-SCF Geometry Optimization. As seen above, the QM/MM-RWFE-SCF geometry optimizations converged after the sequential sampling with the MD simulations for 90 and 21 ns in total for the reactant and the product, respectively, which are much longer than those in previous studies.^{23,25,29,30} During the MD simulations, the protein structures around the binding site underwent large conformational changes, as shown in Figure 8. A loop, L1, approaches the substrate and forms interaction with it in both of the reactant and product states in a similar fashion. Another loop, L2, interacting with the substrate from the other side of L1 also changes its conformation in the reactant state, whereas it does not exhibit large changes in the product state. Consequently, the L2 loop forms an extensive interaction with glucose($+2$) (the number in parentheses is the index of the subsite) in the reactant state, which is absent in the product structure after the geometry optimization as well as in the initial structure of the optimization.

Figure 9 depicts the time evolution of the conformational changes. The movement of the L1 loop almost completes within 10 ns in both the reactant and the product. On the other hand, the L2 loop keeps fluctuating up to 60 ns in the reactant state. At ~ 60 ns, a hydrogen bond between a carboxylate of Asp264 in the QM region and a hydroxyl group of glucose(-2) in the MM one breaks. Then, the L2 loop forms the extensive interaction with glucose($+2$) and becomes stable after breakage of the hydrogen bond. Finally, the optimization of the QM coordinates converges in equilibration for a few tens of nanoseconds. In the case of the product state, the convergence is attained much earlier because the hydrogen bond breaks at a few nanoseconds and the movement of the L2 loop is smaller.

The observations above demonstrate that the present method is capable of determining the optimal QM geometry on an

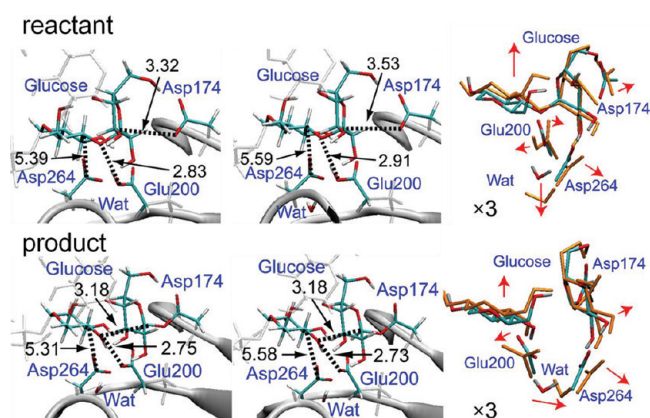


Figure 10. Changes of the QM geometries obtained with the QM/MM free energy optimizations. The left and middle panels depict the initial QM structures determined by the QM/MM potential energy geometry optimization and the free energetically optimized QM ones, respectively. Distances between Asp174:O(6) and glucose:C(41) for the reactant, glucose:O(40) and glucose:C(41) for the product, Glu200:O(12) and glucose:O(40), and Asp264:C(17) and glucose:C(30) are shown (see Figure S1, Supporting Information, for the atom index). The right panels illustrate overall changes of the QM geometries. Structures drawn in colors based on the atom type and in orange indicate the initial ones and the final ones where the structural changes are magnified by a factor of 3 for emphasis.

extensive free energy surface of the MM distribution by following slow conformational relaxation of the protein on a submicrosecond scale. In the present case, one reason for the conformational changes at L1 and L2 in the relaxation are presumably due to a distorted initial conformation by crystal packing of the X-ray crystallographic model since those loops which were found to be flexible³⁵ are in contact with an adjacent protein in the crystal packing. The present free energy geometry optimization method therefore removes properly the possible large distortion of the protein conformation that is hardly detectable by shorter MD simulations.

Furthermore, more notably, the submicrosecond MD searches from the similar initial protein conformations for the reactant and the product found the remarkably different conformations of the L2 loop depending on the catalytic reaction states in the QM region (Figure 8), which cannot be found with an MD conformational search on a less than nanoseconds scale. The observation may imply that the present free energy geometry optimization identified successfully the large and slow conformational change of the protein that couples with the catalytic reaction step and thus plays a role in the enzymatic catalysis, although further examination is necessary for the proposal because of the possibility that those conformations are trapped in local free energy minima.

Optimized Structures of the Catalytic Site. Figure 10 shows comparison between the QM structures optimized by the present method and those by a QM/MM method based on the potential energy surface. In the reactant state, the intramolecular distances between the substrate and three carboxyl groups in the catalytic site, i.e., Asp174, Glu200, and Asp264, increase on the free energy surface (see also Figure S6 in the Supporting Information). The thermal fluctuation of the MM region taken into account therefore relaxes the interaction for the substrate binding. Two negatively charged carboxylates among them, which repel each other, may induce expansion of the binding pocket in thermal fluctuation and thus enhance the relaxation of the interaction.

In the product state, the distance between the substrate and Asp264 increases as well as in the reactant state because of the dissociation of a hydrogen bond described above (Figure 9). On the other hand, intermolecular distances between the groups involved in the catalytic reaction do not undergo large changes. As shown in Figure 10, the atoms that form the glycosidic bond in the reactant state, i.e., O(40) and C(41) (the numbers in parentheses are the atom indices defined in Figure S1, Supporting Information), keep their mutual distance of 3.18 Å in the product state upon the free energy optimization (see also Figure S6, Supporting Information). Moreover, the distance between the glycosidic oxygen atom, O(40), and the O(12) atom of Glu200, which is the proton donor for the dissociation of the glycosidic bond, becomes even shorter than that optimized on the potential energy surface, 2.75 Å \rightarrow 2.73 Å. Their relatively short non-bonding distances compared with those in a nonreactive condition (e.g., 3.59 Å for O \cdots C estimated with the LJ parameters of Amber force field) manifest a strong electronic interaction in the reaction core region. Note that the product state of the reaction step considered in the present study corresponds to an intermediate of the overall enzymatic reaction as mentioned above. Thus, the reaction core region in the product state is still reactive, and the partial bonding character for the seemingly nonbonding interactions among those reaction core atoms remains. The present method is therefore capable of describing such complex electronic interactions quantum mechanically in a thermally fluctuating environment.

ES Potential of the Catalytic Site. Figure 11 shows the ES potential produced by the MM region acting on the QM catalytic site, which plays a role in the enzymatic catalysis. The extensive relaxations of the protein observed in the free energy optimization alter the ES potential largely. Figure 11a displays differences in the mean field ES potentials acting on the QM atoms between the initial and final MM distributions of the sequential sampling, i.e., $\Delta_{ss-opt} V_A = V_A(\text{final step}) - V_A(\text{initial step})$. The initial MM distribution was obtained for the geometry and the charges of the QM molecules determined by the QM/MM potential energy geometry optimization without reweighting and thus corresponds to that used in an approximate scheme for estimation of the QM/MM reaction free energy.^{47,48} Large negative peaks of the ES potential differences around O(18) of Asp264 and C(50) of glucose (−1) are due to the dissociation of a hydrogen bond of Asp264 with a hydroxy group of the neighboring glucose (−2) (see Figure 9). As the absolute values of the ES potential in the QM region are positive because of a negative net charge (−2) of the QM region, the negative differences of the ES potentials indicate decreases of the positive ES potentials. One can also discern large changes of the ES potentials on glucose(+1), which originate from the large movements of loops, L1 and L2 (see Figure 9).

Figure 11b shows differences in the mean field ES potentials between the reactant and the product at the free energetically optimized states, i.e., $\Delta_{\text{reaction}} V_A = V_A(\text{product}) - V_A(\text{reactant})$, which identify important ES interactions for the catalysis. A large decrease and increase of ES potentials on Asp174 and Glu200, respectively, represent ES reorganizations of the protein for changes of protonated states of those groups upon the reaction (see Figure 1). Several large peaks resulting from ES reorganization are also found in glucose(−1). Furthermore, increases of the ES potential are extended over the regions of Asp264 and glucose(+1). In order to examine the origins of the ES potential differences, we evaluated differences in $\Delta_{\text{reaction}} V_A$ between the

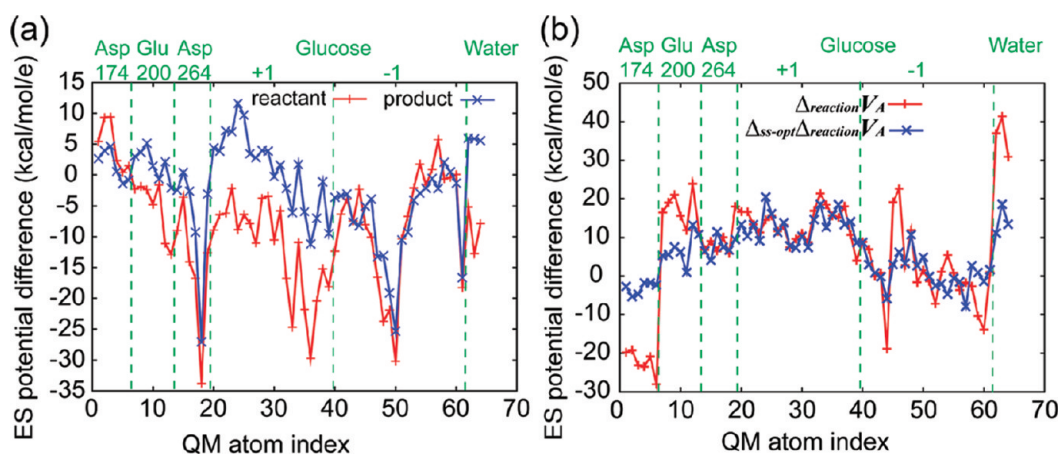


Figure 11. (a) Differences in the mean field ES potentials acting on the QM atoms between the initial and final MM distributions of the sequential sampling, $\Delta_{ss-opt}V_A = V_A(\text{final step}) - V_A(\text{initial step})$. (b) Differences in the mean field ES potentials between the reactant and the product. Red and blue lines indicate differences in the mean field ES potentials between the reactant and the product at the free energetically optimized states, $\Delta_{\text{reaction}}V_A = V_A(\text{product}) - V_A(\text{reactant})$, and differences in $\Delta_{\text{reaction}}V_A$ between the initial and final MM distributions of the sequential sampling, $\Delta_{ss-opt}\Delta_{\text{reaction}}V_A = \Delta_{\text{reaction}}V_A(\text{final step}) - \Delta_{\text{reaction}}V_A(\text{initial step})$, respectively.

initial and final MM distributions of the sequential sampling, i.e., $\Delta_{ss-opt}\Delta_{\text{reaction}}V_A = \Delta_{\text{reaction}}V_A(\text{final step}) - \Delta_{\text{reaction}}V_A(\text{initial step})$, which express ES reorganization due to the conformational changes resulting from the sequential sampling. Unlike other regions, $\Delta_{\text{reaction}}V_A$ of Asp264 and glucose(+1) are close to $\Delta_{ss-opt}\Delta_{\text{reaction}}V_A$, indicating that the ES reorganization comes from the large conformational changes of the L1 and L2 loops. On the other hand, the large $\Delta_{\text{reaction}}V_A$'s of Asp174 are mainly attributed to ES reorganization in a linear response regime because of small $\Delta_{ss-opt}\Delta_{\text{reaction}}V_A$'s which indicate that the ES reorganization is already present in the initial MM distribution of the sequential sampling. It is noteworthy that large positive $\Delta_{ss-opt}\Delta_{\text{reaction}}V_A$'s are found for the reaction core atoms, O(12) (13.2 kcal/mol/e) and H(13) (10.8 kcal/mol/e) of Glu200, and the glycosidic oxygen, O(40) (8.7 kcal/mol/e), which contribute to catalysis of the reaction by stabilization of the developing negative partial charges on those atoms upon the reaction. As the L2 loop is in close proximity to those reaction core atoms, the large movement of L2 accompanied by the reaction is responsible for the generation of the positive ES potential differences.

Comparison with Other QM/MM Free Energy Methods.

The present QM/MM-RWFE-SCF method is based on a combination of theories developed by Yamamoto³⁰ and Yang and co-workers.²⁵ As demonstrated above, the highly improved computational efficiency furnished by the combination allows one to search a QM/MM optimized structure on an unprecedentedly extensive free energy surface. In order to clarify the significant feature of the present method, we compare the present method with other QM/MM methods for the examination of the reaction free energy profile.

Aguilar and co-workers proposed a MM mean field QM/MM method called ASEP/MD.^{28,29} In this method, the MM ES mean field is represented by point charges on grid points around the QM region that are fitted so as to reproduce the mean field acting on the QM region. This contraction of the MM ES mean field reduces the computational cost for time-consuming direct statistical averaging of the ES one electron integral given by eq 11. Theoretically, the description of the QM/MM ES mean field interaction used in the ASEP/MD method based on a form of

eq 11 is more precise than that with the RESP operator introduced in the present study if the mean field is accurately reproduced by the contraction. However, the method suffers from a drawback of inconsistency of the MM thermal distribution. In the case that the QM-MM ES interaction is described in a one electron integral form of eq 11, one needs to compute the time-consuming one electron integrals at each step of a MD trajectory calculation for evaluation of the MM distribution in order to keep the consistency of the MM distribution. The computation of one electron integrals increases drastically the computational time of the MD simulation and thus limits severely the sampling time. In the ASEP/MD method, therefore, the QM-MM ES interaction in the MD simulation is approximately evaluated in a classical Coulombic form with RESP charges of the QM atoms. However, the approximate description of the QM-MM ES interaction in the MD simulation introduces inconsistency with that in the QM/MM geometry optimization in the mean field of MM distribution and in fact led to a slow convergence behavior of the geometry optimization.²⁹ Such a poor convergence might make calculation of the Hessian matrix required for transition state determination and vibrational modes difficult. Furthermore, since the reweighting scheme cannot be applied to the contracted ES mean field, the method requires very frequent updates of the MM distribution during the geometry optimization, which is not suitable for protein systems with slow relaxation (see also below).

Yamamoto developed a mean field QM/MM theory, QM/MM-FE with mean-field embedding,³⁰ on which the mean field ES interaction term of the Fock or KS equation, eq 14, employed in the present study is derived. Although the sequential sampling scheme is also utilized in the method, the MM distribution is not changed during a QM/MM free energy geometry optimization cycle. The iteration scheme simplifies the SCF cycle of the electronic function; the update of the ES mean field by eqs 16 and 17 at each SCF step is not necessary. However, as mentioned above, the unchanged MM distribution is no longer optimal once the electronic wave function (i.e., the charges) and the geometry of the QM region are updated in the QM/MM free energy geometry optimization cycle. Hence, the variational condition for the free energy functional is not satisfied at most steps of the

optimization cycles. The violation of the variational condition causes an arrested minimum search on the free energy surface, leading to slow convergence of the geometry optimization. Figure 6 depicts RMSDs of the first four optimization steps without the reweighting scheme. The RMSD step sizes that result from the optimization searches are considerably small compared with those from the present method. Furthermore, the Hessian matrix calculation by a finite differential method is more difficult. In principle, the MM conformational samples need to be obtained for each of the QM coordinates with small displacements, whereas the reweighting scheme allows one to employ the same samples throughout the calculation as described above.

Yang and co-workers developed the QM/MM-MFEP method,^{24,25} where the reweighting scheme for the QM/MM free energy optimization is introduced. Unlike the other methods described above and the present one, an electronic wave function variational to a free energy functional is not directly solved. Instead, an approximate Hamiltonian function is defined with a reference electronic wave function and its charge response kernel (CRK)^{49,50} that describes linear QM charge response to the MM ES field. Free energy geometry optimization with the sequential sampling is performed on the free energy surface of the approximate Hamiltonian. From a theoretical point of view, the method is advantageous over the mean field approximation since the treatment takes into account the linear response fluctuation of the QM charges in the free energy function, although the effect of the fluctuation of the QM charges is expected to be very minor for the present protein system, as described above. However, computationally, the method includes shortcomings. One is determination of the reference electronic wave function. The change in the ES interaction due to polarization of the QM wave function is represented only by the linear changes of the QM charges from values obtained for the reference electronic wave functions. Thus, the reference electronic wave function needs to be close to the optimal one in order to describe accurately the QM polarization by the linear response approximation. However, no reasonable way to determine such a good reference electronic wave function has been proposed. It is suggested that the present mean field method would provide a good reference wave function for the linear treatment of the QM polarization. Another drawback is that the method is not variational with respect to the electronic wave function. Hence, evaluation of the free energy gradient requires the calculation of coupled perturbed equations, which reduces computational efficiency. Furthermore, MD trajectory calculations for MM conformational samplings were very limited in the studies reported in refs 24 and 25 (approximately 100 ps for each sequential sampling). As revealed above, however, the reweighting scheme is very sensitive to the MM distribution and leads easily to the serious ill behavior of the reweighted distribution for an insufficiently sampled MM distribution. Careful examination of the reweighted MM distribution as carried out in the present study is suggested when the reweighting scheme is used.

Finally, QM/MM-MD simulations with empirical QM methods such as EVB,¹⁹ MCMM,^{20,21} and DFTB^{17,18} are mentioned. Those empirical methods are furnished with computational efficiency enough to carry out MD simulations for a relatively long time with accuracy attained through parametrizations of empirical Hamiltonians so as to reproduce energies and forces of high-level ab initio QM or QM/MM calculations. The direct MD simulations with QM/MM Hamiltonians allow one to sample

thermal distributions of the QM coordinates which cannot be obtained by the QM/MM geometry optimization methods described above. However, in general, accurate parametrizations of the Hamiltonian become difficult for complex reactions where the QM systems are strongly correlated with large conformational changes of the MM surroundings. It is therefore suggested that the present method is suitably employed to obtain the reference energies and geometries for the parametrization because the method provides more accurate descriptions of electronic wave function, geometry, and normal modes at the special states of reaction than conventionally used ones, i.e., gas phase QM methods or QM/MM ones based on potential energy surfaces.

CONCLUSION

We developed a QM/MM free energy optimization method by combining a mean field QM/MM theory³⁰ with a reweighting technique for the MM ensemble averaging.²⁵ This QM/MM-RWFE-SCF method features applicability for enzymatic reactions that involve extensive, heterogeneous, and slow thermal relaxation of the protein. Its high efficiency of computational scheme and precise description for long-range ES interaction using the Ewald summation technique enable one to explore an enzymatic reaction on an extensive free energy surface of the protein conformation. We demonstrated free energy geometry optimizations of the reactive substrate following global and non-linear protein conformational changes of the α -amylase protein on a time scale reaching the submicrosecond level. The free energy geometry optimizations revealed that a loop adjacent to the catalytic site forms in significantly different conformations in the reactant and the product, respectively, and produces a catalytic ES field for the enzymatic reaction. The method now opens the way for theoretical examination of a proposal on enzymatic catalysis by slow protein dynamics, which was under debate recently.^{51,52}

ASSOCIATED CONTENT

S Supporting Information. Index for atoms in the QM site, differences between Ewald ES potentials and Ewald ES potentials without interaction between the QM regions in real and reciprocal space, differences between Ewald ES forces and Ewald ES forces without interaction between the QM regions in real and reciprocal space, convergence of Ewald ES potentials with respect to the number of k vectors of reciprocal space, correlation between QM charges evaluated using the QM/MM-RWFESCF method and those using the average without the mean field approximation for 2000 conformations in the last MM distribution of the sequential sampling, and changes of distances between atoms in the QM sites during the sequential samplings. This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*Tel.: +81-75-753-4006. Fax: +81-75-753-4000. E-mail: hayashig@kuchem.kyoto-u.ac.jp.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

The study was supported by research fellowships for young scientist from the Japan Society for the Promotion of Science (JSPS) to T.K., by a Grant-in-Aid for Scientific Research on Priority Areas (18074004) and that on Innovative Areas (23107717) from the Ministry of Education, Culture, Sports, Science, and Technology, Japan; by a Grant-in-Aid for Scientific Research from JSPS (23700580); by Research and Development of the Next-Generation Integrated Simulation of Living Matter; and the Global COE program "International Center for Integrated Research and Advanced Education in Materials Science". The molecular images were created with VMD.⁵³

REFERENCES

- (1) Fersht, A. *Enzyme Structure and Mechanism*, 2nd ed; W. H. Freeman and Company: New York, 1985.
- (2) Suckling, C. *Enzyme Chemistry: Impact and application*, 2nd ed; Chapman and Hall: London, U.K., 1990.
- (3) Warshel, A.; Levvit, M. *J. Mol. Biol.* **1976**, *103*, 227–229.
- (4) Field, M. J.; Bash, P. A.; Karplus, M. *J. Comput. Chem.* **1990**, *6*, 700–733.
- (5) Gao, J. *Acc. Chem. Res.* **1996**, *29*, 298–305.
- (6) Svensoon, M.; Humbel, S.; Froese, R. D. J.; Matsubara, T.; Sieber, S.; Morokuma, K. *J. Phys. Chem.* **1996**, *100*, 19357–19363.
- (7) Monard, G.; Merz, K. M., Jr. *Acc. Chem. Res.* **1999**, *32*, 904–911.
- (8) Hess, B. *Phys. Rev. E* **2000**, *62*, 8438–8448.
- (9) Hess, B. *Phys. Rev. E* **2002**, *65*, 031910.
- (10) Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E. *Curr. Opin. Struct. Biol.* **2009**, *19*, 120–127.
- (11) Nishihara, Y.; Kato, S.; Hayashi, S. *Biophys. J.* **2010**, *98*, 1649–1657.
- (12) Jimenez, R.; Fleming, G. R.; Kumar, P. V.; Maroncelli, M. *Nature* **1994**, *369*, 471–473.
- (13) Devi-Kesavan, L. S.; Gao, J. *J. Am. Chem. Soc.* **2003**, *125*, 1532–1540.
- (14) Ridder, L.; Rietjens, I. M. C. M.; Vervoort, J.; Mulholland, A. J. *J. Am. Chem. Soc.* **2002**, *124*, 9926–9936.
- (15) Ruiz-Pernía, J. J.; Silla, E.; Tuñón, I. *J. Phys. Chem. B.* **2006**, *110*, 20686–20692.
- (16) Lameira, J.; Alves, C. N.; Moliner, V.; Martí, S.; Castillo, R.; Tuñón, I. *J. Phys. Chem. B.* **2010**, *114*, 7029–7036.
- (17) Han, W.; Elstner, M.; Jalkanen, K. J.; Frauenheim, T.; Suhai, S. *Int. J. Quantum Chem.* **2000**, *78*, 459–479.
- (18) Cui, Q.; Elstner, M.; Frauenheim, T.; Kaxiras, E.; Karplus, M. *J. Phys. Chem. B.* **2001**, *105*, 569–585.
- (19) Warshel, A.; Weiss, R. M. *J. Am. Chem. Soc.* **1980**, *102*, 6218–6226.
- (20) Kim, Y.; Corchado, J. C.; Villà, J.; Xing, J.; Truhlar, D. G. *J. Chem. Phys.* **2000**, *112*, 2718–2735.
- (21) Higashi, M.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 790–803.
- (22) Okuyama-Yoshida, N.; Nagaoka, M.; Yamabe, T. *Int. J. Quantum Chem.* **1998**, *70*, 95–103.
- (23) Okuyama-Yoshida, N.; Kataoka, K.; Nagaoka, M.; Yamabe, T. *J. Chem. Phys.* **2000**, *113*, 3519–3524.
- (24) Hu, H.; Lu, Z.; Yang, W. *J. Chem. Theory Comput.* **2007**, *3*, 390–406.
- (25) Hu, H.; Lu, Z.; Parks, J. M.; Burger, S. K.; Yang, W. *J. Chem. Phys.* **2008**, *128*, 034105.
- (26) Higashi, M.; Hayashi, S.; Kato, S. *Chem. Phys. Lett.* **2007**, *437*, 293–297.
- (27) Higashi, M.; Hayashi, S.; Kato, S. *J. Chem. Phys.* **2007**, *126*, 144503.
- (28) Sánchez, M. L.; Aguilar, M. A.; Olivares del Valle, F. *J. Comput. Chem.* **1997**, *18*, 313–322.
- (29) Galván, I. F.; Sánchez, M. L.; Martín, M. E.; Olivares del Valle, F. J.; Aguilar, M. A. *J. Chem. Phys.* **2003**, *118*, 255–263.
- (30) Yamamoto, T. *J. Chem. Phys.* **2008**, *129*, 244104.
- (31) Siddiqui, K. S.; Cavicchioli, R. *Annu. Rev. Biochem.* **2006**, *75*, 403–433.
- (32) D'Amico, S.; Gerday, C.; Feller, G. *J. Biol. Chem.* **2001**, *276*, 25791–25796.
- (33) D'Amico, S.; Gerday, C.; Feller, G. *J. Biol. Chem.* **2002**, *277*, 46110–46115.
- (34) D'Amico, S.; Marx, J.-C.; Gerday, C.; Feller, G. *J. Biol. Chem.* **2003**, *278*, 7891–7896.
- (35) Kosugi, T.; Hayashi, S. *Chem. Phys. Lett.* **2011**, *501*, 517–522.
- (36) Qian, M.; Nahoum, V.; Bonicel, J.; Bischoff, H.; Henrissat, B.; Payan, T. *Biochemistry* **2001**, *40*, 7700–7709.
- (37) Hayashi, S.; Ohmine, I. *J. Phys. Chem. B.* **2000**, *104*, 10678–10691.
- (38) Essman, V.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593.
- (39) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani, F.; Wu, X.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathews, D. H.; Schafmeister, C.; Ross, W. S.; Kollman, P. A. *AMBER 9*; University of California: San Francisco, CA, 2006.
- (40) Aghajari, N.; Roth, M.; Haser, R. *Biochemistry* **2002**, *41*, 4273–4280.
- (41) Kirschner, K. N.; Woods, R. J. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10541–10545.
- (42) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (43) Schmidt, M. W.; Baldrige, 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.
- (44) Kaukonen, M.; Söderhjelm, P.; Heimdal, J.; Ryde, U. *J. Chem. Theory Comput.* **2008**, *4*, 985–1001.
- (45) Sánchez, M. L.; Martín, M. E.; Galván, I. F.; Olivares del Valle, F. J.; Aguilar, M. A. *J. Phys. Chem. B.* **2002**, *106*, 4813–4817.
- (46) Galván, I. F.; Martín, M. E.; Aguilar, M. A.; Ruiz-López, M. F. *J. Chem. Phys.* **2006**, *124*, 214504.
- (47) Zhang, Y.; Liu, H.; Yang, W. *J. Chem. Phys.* **2000**, *112*, 3483–3492.
- (48) Ishida, T.; Kato, S. *J. Am. Chem. Soc.* **2003**, *125*, 12035–12048.
- (49) Morita, A.; Kato, S. *J. Am. Chem. Soc.* **1997**, *119*, 4021–4032.
- (50) Lu, Z.; Yang, W. *J. Chem. Phys.* **2004**, *121*, 89–100.
- (51) Nagel, Z. D.; Klinman, J. P. *Nat. Chem. Biol.* **2009**, *5*, 543–550.
- (52) Kamerlin, S. C. L.; Warshel, A. *Proteins* **2010**, *78*, 1339–1375.
- (53) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graph.* **1996**, *14*, 33–38.

Understanding the Sequence Preference of Recurrent RNA Building Blocks Using Quantum Chemistry: The Intrastrand RNA Dinucleotide Platform

Arnošt Mládek,^{*,†} Judit E. Šponer,^{†,‡} Petr Kulhánek,^{‡,§} Xiang-Jun Lu,^{||} Wilma K. Olson,[⊥] and Jiří Šponer^{*,†,‡}

[†]Institute of Biophysics, Academy of Sciences of the Czech Republic, Královopolská 135, 612 65 Brno, Czech Republic

[‡]CEITEC - Central European Institute of Technology, Masaryk University, Campus Bohunice, Kamenice 5, 625 00 Brno, Czech Republic

[§]National Centre for Biomolecular Research, Faculty of Science, Masaryk University, 611 37 Brno, Czech Republic

^{||}Department of Biological Sciences, Columbia University, New York, New York 10027, United States

[⊥]Department of Chemistry and Chemical Biology, BioMaPS Institute for Quantitative Biology, Rutgers—The State University of New Jersey, Piscataway, New Jersey 08854, United States

S Supporting Information

ABSTRACT: Folded RNA molecules are shaped by an astonishing variety of highly conserved noncanonical molecular interactions and backbone topologies. The dinucleotide platform is a widespread recurrent RNA modular building submotif formed by the side-by-side pairing of bases from two consecutive nucleotides within a single strand, with highly specific sequence preferences. This unique arrangement of bases is cemented by an intricate network of noncanonical hydrogen bonds and facilitated by a distinctive backbone topology. The present study investigates the gas-phase intrinsic stabilities of the three most common RNA dinucleotide platforms—5'-GpU-3', ApA, and UpC—via state-of-the-art quantum-chemical (QM) techniques. The mean stability of base–base interactions decreases with sequence in the order GpU > ApA > UpC. Bader's atoms-in-molecules analysis reveals that the N2(G)···O4(U) hydrogen bond of the GpU platform is stronger than the corresponding hydrogen bonds in the other two platforms. The mixed-pucker sugar–phosphate backbone conformation found in most GpU platforms, in which the 5'-ribose sugar (G) is in the C2'-endo form and the 3'-sugar (U) in the C3'-endo form, is intrinsically more stable than the standard A-RNA backbone arrangement, partially as a result of a favorable O2'···O2P intraplatform interaction. Our results thus validate the hypothesis of Lu et al. (Lu, X.-J.; et al. *Nucleic Acids Res.* **2010**, *38*, 4868–4876) that the superior stability of GpU platforms is partially mediated by the strong O2'···O2P hydrogen bond. In contrast, ApA and especially UpC platform-compatible backbone conformations are rather diverse and do not display any characteristic structural features. The average stabilities of ApA and UpC derived backbone conformers are also lower than those of GpU platforms. Thus, the observed structural and evolutionary patterns of the dinucleotide platforms can be accounted for, to a large extent, by their intrinsic properties, as described by modern QM calculations. In contrast, we show that the dinucleotide platform is not properly described in the course of atomistic explicit-solvent simulations. Our work also gives methodological insights into QM calculations of experimental RNA backbone geometries. Such calculations are inherently complicated by rather large data and refinement uncertainties in the available RNA experimental structures, which often preclude reliable energy computations.

INTRODUCTION

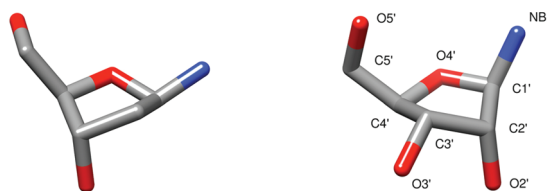
Nucleic acids (NA) are polymeric biomacromolecules that play vital roles in cellular life. The primary and most essential function of 2'-deoxyribonucleic acid (DNA) is to preserve the genetic information in the cell. Conversely the pool of functions for which ribonucleic acid (RNA) is accountable is much larger and still not exhaustively explored and fully comprehended. While less than 2% of the human genomic DNA directly encodes protein sequences, over 80% of the genome is actually transcribed into RNA. Thus, the vast majority of the genome encodes nonprotein coding RNAs (ncRNAs) with numerous known as well as hitherto unknown functions.

The ability of RNA molecules to execute miscellaneous tasks has its origin in the tremendous variability of complexly organized structures, which are made possible by the 2'-hydroxyl

group attached to the C2' atom of the sugar moiety (Figure 1). The easily accessible 2'-hydroxyl group features a hydrogen bond (H bond) with both acceptor and donor capabilities, which allow it to interact and stabilize complex tertiary structures and modular motifs indispensable to RNA organization and inherently inaccessible to DNA. The 2'-hydroxyl group also represents one of the crucial components of the so-called sugar edge of a ribonucleotide. RNA molecules create a wide variety of base pairs by systematically combining the three edges of the constituent ribonucleotides, i.e., the sugar edge, the Watson–Crick edge, and the Hoogsteen edge.^{1–3} Such interactions define the shape and conservation patterns of folded, nonhelical regions of RNA.⁴

Received: October 7, 2011

Published: December 08, 2011



DNA sugar: C2'-endo (B DNA) RNA sugar: C3'-endo (A RNA)

Figure 1. The sugar moieties of DNA (2'-deoxyribose, left) and RNA (ribose, right). The 2'-hydroxyl group of ribose is a powerful donor and acceptor of hydrogen bonds. The illustrated puckering of the sugar rings corresponds to the forms that prevail in B-DNA (C2'-endo, left) and A-RNA (C3'-endo, right) structures. Oxygen is depicted in red, carbon in gray, and nitrogen in blue. The "NB" label denotes the nitrogen atom, either N9 (purine) or N1 (pyrimidine), via which the nucleobase is linked to the anomeric C1' atom of the sugar. Hydrogens are omitted for the sake of clarity.

The complexity of RNA interactions, however, is even greater than that suggested by the combination of three nucleotide edges used in standard RNA base-pairing classifications.^{1–4} For example, many base–base, base–sugar, and sugar–sugar H bonds occur in concert with highly conserved base–phosphate H bonds, a classification of which has been proposed by combining RNA structural bioinformatics and QM approaches.^{5,6} The preference of the 2'-hydroxyl group for particular H-bond acceptors—such as the phosphodiester bridging oxygens (O3'(n) and O5'(n + 1), where "n" denotes the residue number in the 5' → 3' direction of the RNA chain), the anionic phosphate oxygens (O1P and O2P of the (n + 1)-th residue), or the adjacent sugar ring oxygen (O4'(n + 1))—is strongly modulated by the sugar–phosphate backbone conformation and vice versa. Thus, the 2'-hydroxyl group affects the conformation of both the backbone and the sugar ring. This coupling of structural variables makes description of the RNA sugar–phosphate backbone potential energy surface a more challenging task than that of double-helical DNA. Whereas QM studies of base stacking and base pairing are relatively easy and thus abundant in the literature, including investigations specifically devoted to RNA interactions,^{7–17} there are rather few QM studies that deal with the stabilizing features and basic conformational properties of the sugar–phosphate backbone.^{18–30} Moreover, most of the published studies deal with the DNA backbone, partially because the 2'-OH of ribose complicates QM computations; i.e., the hydroxyl group tends to form spurious biochemically irrelevant H bonds in model computations. At the same time, QM studies of the NA backbone are needed, since adequate description of the sugar–phosphate backbone is a notorious weakness of molecular simulation force fields. Subtle imbalances in the description of the nucleic acid backbone can lead to the entire degradation of simulated nucleic acid systems.^{30–32} In addition, description of the backbone is challenging experimentally. For example, the sugar atoms are much less visible than the nucleobases and phosphates in crystallographically derived electron densities. Inherent flexibility of the backbone often precludes unambiguous refinement of the nucleotide conformation from such data.

An understanding of the principles of RNA folding is essential to resolution of the mechanisms that underlie the multitude of functions that RNA molecules execute. RNA structures often assemble in a modular fashion and make use of evolutionarily conserved sequences and three-dimensional (3D) patterns to

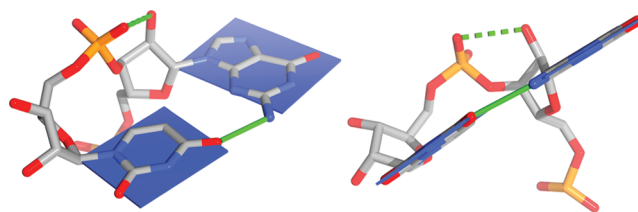


Figure 2. Two views of a dinucleotide platform submotif (here 5'-GpU-3'). The blue slabs depict the planes of the adjacent nucleobases. The approximate coplanarity of the bases is a distinctive feature of all dinucleotide platforms. The green solid lines denote the interbase side-by-side pairing, and the green dashed line symbolizes the O2'...O2P H bond typical of GpU.³³

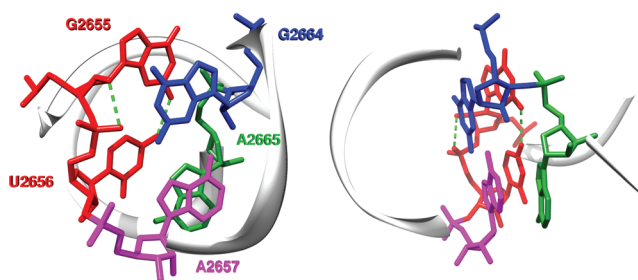


Figure 3. Image of the sarcin/ricin domain in the ultra-high-resolution (1.00 Å) structure of *Escherichia coli* 23S rRNA³⁶ (PDB ID: 3dvz). The highly conserved asymmetric GpUpA/GpA miniduplex is depicted at the atomic level using a stick representation. Left, view along the helical axis of the duplex; right, side view. The GpU dinucleotide platform (G2655pU2656) is colored in red, and its two key N2(G)···O4(U) and O2'(G)···O2P(U) H bonds are depicted by green dashed lines. For the sake of clarity, the remainder of the domain is represented by a backbone trace, and numerous other H bonds stabilizing the GpUpA/GpA miniduplex are not depicted (see Figure 2 of ref 33 for details). Among them, interactions of A2665 (green) from the opposite strand with the GpU platform (red) form a characteristic in-plane nucleobase triad. The GpU···A2665 in-plane arrangement is stabilized by U···A interbase H-bonding (trans Watson–Crick/Hoogsteen pattern^{1,2}) and G···A base–phosphate interaction (4BPh class,⁵ see below). Note that the G2655 base of the GpU platform is bulged out of the noncanonical RNA double helix. The remaining two bases, G2664 (blue) and A2657 (magenta), form a sheared GA base pair that stacks on the triad and completes the miniduplex motif.

perform various tasks. Therefore, knowledge of the stabilities of the structural modules provides a rationale for their evolutionary conservation and may elucidate why selected nucleotides at the primary informational level are critical for module performance.

The present paper investigates one of the most interesting RNA 3D structural submotifs known as a dinucleotide platform. By submotif, we mean a characteristic noncanonical RNA 3D element that does not fold independently and that requires auxiliary structural elements to form an autonomous 3D building block. The intrastrand dinucleotide platforms are modules formed by two adjacent nucleotides 5'-XpY-3' with side-by-side XY paired nucleobases. In other words, the unique backbone topology of the dinucleotide platform places the two consecutive bases in a common plane (Figure 2). Examples of dinucleotide platforms have been experimentally identified in a variety of atomic-resolution structures, including (i) GpU platforms in the

complex of a small fragment of *Escherichia coli* 23S rRNA with the ribosomal L11 protein,³⁴ the sarcin/ricin domain of the large ribosomal subunit (Figure 3),^{35,36} the hammerhead ribozyme,³⁷ and other parts of the large ribosomal subunit;³⁸ (ii) ApA platforms in the P4–P6 domain of a group I intron³⁹ and the large ribosomal subunit;³⁸ and (iii) UpC platforms in the genomic ribozyme precursor of the hepatitis delta virus⁴⁰ and the cysteinyl–tRNA synthetase binary complex with tRNA^{Cys},⁴¹ etc. The resolution of the experimental structures ranges between 0.97 and 2.40 Å. Note that the majority of the listed experimental structures were determined at a resolution poorer than 1.5 Å. Despite being satisfactory for many purposes, such resolution does not guarantee unambiguous determination of the fine structural details of the sugar–phosphate backbone. QM computations are highly sensitive to any such uncertainties in the conformations of individual platforms.

The GpU platforms constitute over half of the identified dinucleotide platforms.³³ Analysis of the experimental geometries strongly suggests that their widespread presence may reflect a favorable interaction between the 2'-hydroxyl group of the 5'-residue (G) and the anionic phosphate group oxygen (O2P) of the following 3'-nucleotide (U) (cf. Figure 2). The apparent intrinsic stability of the GpU platform correlates with the highly conserved and naturally stiff sugar–phosphate backbone conformation,³³ which enables formation of the O2'···O2P H-bond.

Our aim is to gain a better understanding of the clear prevalence of the GpU platform compared to the ApA and UpC platforms (the second and third most frequent platforms) in known high-resolution structures by means of modern QM calculations. More specifically, we want to determine whether the frequency of occurrence of the different dinucleotide platforms stems from their intrinsic stabilities and electronic structures. Indeed, we show that the GpU platform is intrinsically more stable than the ApA and UpC analogs and that this stability is captured by advanced electronic structure computations even in a small model system in the gas phase. The predominance of the GpU and ApA platforms in RNA molecules is clear-cut.³³ The frequency of occurrence of the third most frequent platform (UpC), however, does not differ noticeably from that of ApC, CpA, and GpG (see Table 1 of ref 33). As we are primarily interested in platform stabilization characteristics rather than stability assessments of each possible platform type, we restrict the present calculations to only one of the less frequent platforms, UpC. The number of examples of UpC platforms in better-resolved structures also exceeds the numbers for the excluded platforms (see below). As shown below, even for the UpC platform, the available experimental structural data preclude reliable QM analysis (see Results and Discussion).

MODEL SYSTEMS

The different atomic compositions of the various dinucleotide platforms rule out direct comparison of their stabilities in terms of the total electronic energies. Thus, we divide the dinucleotide platform into two parts: the common sugar–phosphate backbone fragment, which can be easily compared for all studied dinucleotides, and the adjacent nucleobases alone. The stabilization contribution due to interbase interaction is estimated via a standard interaction energy computation. Note that, unlike the electronic energy, the interaction energy is a size-independent quantity and thus well suited for comparison of base···base stabilization contributions. Hence, it is possible to compare the

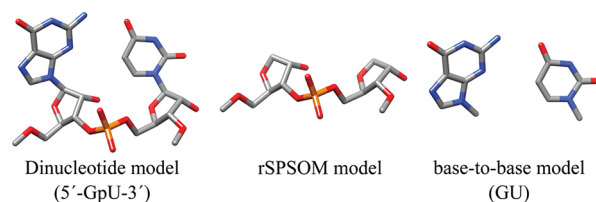


Figure 4. Model systems used to treat platform energetics: (left) full dinucleotide platform (here 5'-GpU-3'); (center) rSPSOM; (right) base-to-base (here GU). Phosphorus is depicted in orange, oxygen in red, nitrogen in blue, and carbon in gray. Hydrogens are omitted for the sake of clarity. The 5' → 3' progression of the RNA chain is from left to right.

interaction energies of base pairs of arbitrary atomic compositions directly. It, however, should be stressed that this partition neglects the potential coupling effects between the two parts of the dinucleotide. In particular, the ApA platform might be stabilized via the O2'···N7(*n* + 1) interaction, which is lost upon division of the system into subsystems.

The selection of dinucleotide platforms is based on a non-redundant data set of experimentally determined structures of 2.5 Å or better resolution characterized using the 3DNA software package^{42,43} (see the Supporting Information for ref 33, Table S2). The set contains a total of 72 dinucleotide platforms, including 43 5'-GpU-3', 15 ApA, 6 UpC, 2 ApC, and 2 CpA platforms, but only single occurrences of CpC, GpA, GpG, and UpA platforms. Of the 43 listed GpU platforms we have selected all (12) platforms resolved at 1.9 Å or better resolution and 17 random platforms from the remaining examples. We have further supplemented the GpU set with three additional platforms, not included in the above data set, from recent ultra-high-resolution structures of the sarcin/ricin domain from *Escherichia coli* 23S rRNA³⁶ (PDB IDs: 3dvz, 3dw4, and 3dw6), corresponding to a total of 32 (12 + 17 + 3) GpU platforms. We also analyzed all but one of the ApA platforms, which is a DNA dinucleotide platform, and five of the six UpC platforms. Although some of the 43 detected GpU platforms were not taken into account, particularly the less well-resolved ones, the principal limitation of the current study is the deficiency in the number of ApA and UpC structures. Therefore, the somewhat arbitrary selection of the 17 poorer resolution GpU platforms has no effect on the conclusions of this study.

The initial geometries of the GpU, ApA, and UpC intrastrand dinucleotide platforms were extracted from the relevant dinucleotide steps in known X-ray structures (Supporting Information, Table S1) and subsequently subjected to a manual two-step chemical modification. First, the chain was terminated with methyl groups, and the appropriate atoms were saturated with hydrogens. Specifically, the phosphate group of the *n*th residue was replaced with a methoxy group (–O–CH₃), and the O3' of the (*n* + 1)th nucleotide was capped with a methyl group (–CH₃). The presence of the methyl groups, rather than hydrogen atoms, at the ends of the backbone chain normally precludes the formation of spurious intramolecular H bonds that may bias the energetics. The 2'-hydroxyl groups of both β-D-ribose groups were initially oriented so that the C3'–C2'–O2'–H2' dihedral angle was equal to 0°, and the C_{Met}–O5'/O3' bond length of both 5' and 3' ends was set to 1.4 Å. Our unpublished data show that this arbitrary initial orientation guarantees the most energetically favorable orientation after gradient optimization. The initial

Table 1. Comparison of the RNA Sugar-Phosphate Backbone Torsion Angles (Degrees) of the Four Idealized Conformational Categories Found in Dinucleotide Platform Structures with Those in Standard A-RNA Steps^a

conformational class ^c	label ^d	average sugar–phosphate torsion angles ^b							
		γ^e	δ	ϵ	ζ	$\alpha + 1$	$\beta + 1$	$\gamma + 1$	$\delta + 1$
I	&a	56	82	−169	−95	−64	−178	51	82
II	#a	164	148	−168	146	−71	151	42	85
III	0a	53	149	−137	139	−75	158	48	84
IV	4g	48	148	−103	165	−155	165	49	83
A-RNA	1a	54	80	−150	−73	−65	173	54	80

^a For the backbone torsion angles nomenclature, see Figure 5. ^b Average backbone torsions found by Richardson et al. (categories I–IV) and Schneider et al. (A-RNA). For more details see refs 44 and 45. ^c Conformational classes based on distinguishing torsional features of the sugar–phosphate–sugar unit: I, A-like homogeneous C3'-endo puckered sugars, *gauche*[−], *gauche*[−] phosphodiester; II and III, mixed C2'-endo/C3'-endo puckering, *trans,gauche*[−] phosphodiester; IV, mixed C2'-endo/C3'-endo puckering, *trans,trans* phosphodiester. ^d Respective RNA family labels according to the Richardson et al. nomenclature.⁴⁴ ^e γ values for classes I–IV taken respectively from appropriate experimental examples (PDB ID/Nucleotide/Chain): 1hq1/163/B;⁴⁶ 1s72/1371,⁴⁷ 1s72/265,⁴⁷ and 1hr2/226/A.⁴⁸ The idealized A-RNA structure is periodic, and thus γ is identical to $\gamma + 1$.

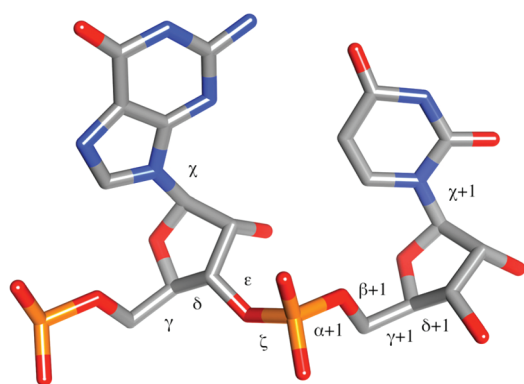


Figure 5. Standard labeling of the sugar–phosphate backbone (α – ζ) and glycosidic (χ) torsion angles of a dinucleotide unit ($5'$ -GpU- $3'$). The “+1” next to a Greek letter denotes a respective backbone torsion of the succeeding nucleotide in the $5' \rightarrow 3'$ direction (from left to right in this figure).

values of the C–H, N–H, and O–H distances were set to 1.09 Å, 1.00 Å, and 1.00 Å, respectively.

Figure 4 shows the three model systems used in the present study. The left image shows the full dinucleotide platform model with methyl-capped ends, while the center image shows a model system, abbreviated as rSPSOM (ribonucleic version of the Sugar–Phosphate–Sugar model with capping -O-Methyl groups), that mimics the backbone segment of the dinucleotide. The rSPSOM model originates from the SPSOM model system, which has been extensively used to study the DNA backbone.^{18,19} The right-most model captures the base···base interaction that is missing from the rSPSOM representation but found in the full dinucleotide model.

The rSPSOM model system (Figure 4, center) has been used to study the intrinsic stability of the platform backbone. The starting structure of the sugar–phosphate fragment in the reduced backbone model is identical to that in the full starting dinucleotide. The starting structures for the base-pair calculations were taken from optimized (i.e., partially relaxed, see below) dinucleotide model geometries.

We have studied two sets of structures: (i) structures taken directly from experiments and (ii) structures based on the idealized backbone conformations⁴⁴ of those backbone families occurring in the dinucleotide platforms. We utilized the

Suitename 0.3 (categorization) Kinamage and Dangle 0.63 (backbone torsion angle determination) software to group the extracted rSPSOMs of the dinucleotide platforms into known conformational categories—see <http://kinamage.biochem.duke.edu>.⁴⁴ Although the conformations of some of the backbone units could not be assigned (13 instances), the majority of structures fell into one of four known categories (out of 46) labeled herein using Roman numerals I–IV: one with the homogeneous C3'-endo sugar puckering and *gauche*[−], *gauche*[−] phosphodiester conformation typical of A RNA (I, one instance), two with mixed C2'-endo/C3'-endo sugar puckering and concomitant rearrangement of the phosphodiester linkage to a *trans,gauche*[−] state (categories II and III with 24 and 6 instances, respectively), and the last with the same mixed puckering and an all-*trans* phosphodiester arrangement (IV, 7 instances). Note that each of the four I–IV categories coincides with a respective RNA backbone family determined in a clustering analysis performed by Richardson et al., see ref 44 and Table 1. The computations on the idealized structures are based on these four groupings, i.e., generated from the average backbone torsions listed in Table 1. Since the conformational groupings are based on a sugar-to-sugar unit (defined by the $5' \rightarrow 3'$ sequence of torsions starting with δ for the first nucleotide and terminating with δ for the last), the γ torsion of the first nucleotide in the rSPSOM model was taken from an appropriate experimental structure (Table 1). The commonly used nomenclature of the six sugar–phosphate backbone torsion angles (labeled as α , β , γ , δ , ϵ , and ζ) and the glycosidic torsion (χ) describing the relative orientation of a nucleobase with respect to the attached sugar ring is given in Figure 5.

The torsion angles of the canonical A-RNA structure used in this paper differ slightly, at most by 2° (Table 1), from those of the reference state of the same name used in the backbone classification software; i.e., we use the older data of Schneider et al.,⁴⁵ which have no effect on the energetics. In view of the periodicity of the A-RNA conformational substate, the γ torsion of the first nucleotide was set to that of the second nucleotide (54°). The values of the quasi- β torsion at the $5'$ end (the quasi prefix indicates that the actual β torsion, defined as P–O5'–C5'–C4', differs from the quasi- β , where the phosphate group is replaced by a methoxy group) and the quasi- $\epsilon + 1$ torsion at the $3'$ end were similarly adjusted to coincide with the β and $\epsilon + 1$ dihedrals in the sugar-to-sugar unit. The main reason why we need to define the quasi torsions is the minimization protocol

issue discussed to more depth in the Supporting Information in the paragraph on A-RNA optimization constraints.

Different sequences presumably place distinct restrictions on the sugar–phosphate backbone geometry in order to form a platform structure in which adjacent bases interact via side-by-side hydrogen bonding. Hence, in additional calculations, we combined the three studied base sequences (GU, AA, UC) with the four identified conformational classes (I–IV), in order to find compatible combinations capable of forming a platform submotif. The fusion was performed by attachment of the selected nucleobases to the C1' atoms of the sugar residues of the idealized optimized rSPSOM model system for each conformational type. The two glycosidic torsions, i.e., χ and $\chi + 1$, were adjusted manually so that the resulting dinucleotide resembled a platform-like geometry. The arranged dinucleotide system was then subjected to a constrained geometry optimization with backbone torsions kept frozen at the initial values, i.e., at the respective class averages listed in Table 1. The compatibility of the base composition and backbone family to form a dinucleotide platform was assessed visually.

COMPUTATIONAL METHODS

The geometries of the dinucleotide and rSPSOM model systems were preoptimized with constrained backbone torsions using the hybrid meta-GGA Minnesota M06 functional⁴⁹ and the 6-31+G(d,p) basis set. The systems were then reoptimized with all previously applied backbone constraints at a higher level using the dispersion-corrected DFT-D approach. We used the meta-GGA TPSS functional⁵⁰ with an entirely local exchange–correlation description augmented with Jurečka's empirical dispersion B-0.96-27 type term (the abbreviation TPSS-D thus marks a particular form of DFT-D method)⁵¹ and combined with the 6-31++G(3df,3pd) set of atomic orbitals, hereafter labeled LP (according to the “Large Pople's basis set”). As we have previously shown,¹⁸ to keep the systems in biologically meaningful conformations, we had to fix the backbone torsions at their experimentally determined values via application of constraints on all of the backbone dihedrals (from γ up to $\delta + 1$ following the 5' → 3' direction). The same set of angles as listed in Table 1 was constrained for experimental as well as idealized structures. In the case of dinucleotide platforms, the two additional glycosidic angles (Figure 5), χ and $\chi + 1$, were constrained at their initial values as well. Some additional constraints on the quasi-torsions were applied in relaxation of canonical A-RNA structure. As these calculations are not essential for our study, the A-RNA constraints are described in the Supporting Information.

The TPSS-D/LP gradients were calculated with Turbomole 5.10⁵² using the resolution of identity (RI) approximation.^{53–55} The empirical dispersion corrections were obtained with an in-house Fortran code and then added to the pure DFT energy upon the gradient calculation run. To take advantage of the efficient and robust optimization algorithms of the Gaussian 03 software package⁵⁶ and the superior scalability of the Turbomole code, we developed a scheme whereby the electronic energy gradients calculated by Turbomole are passed to Gaussian 03 to execute the energetically downhill geometry alteration. The modified geometry is then passed back to Turbomole and serves as a new input structure for the next cycle. This iterative procedure repeats until convergence criteria imposed on the energy and the density matrix are met.

The single-point energies of the rSPSOM model systems were calculated at the RI-MP2/aug-cc-pVDZ and aug-cc-pVTZ^{57,58} levels of theory. We also estimated the energies at the complete basis set limit (RI-MP2/CBS) according to the Halkier et al. extrapolation scheme.^{59,60} The extrapolation to CBS effectively eliminates intramolecular basis set superposition (BSSE) and incompleteness (BSIE) errors, both of which bias the results. Our preceding experience indicates that the aug-cc-pVDZ → aug-cc-pVTZ (D→T) based extrapolation provides results that more likely approach the MP2/aug-cc-pVQZ level rather than true MP2/CBS behavior. Thus, some residual BSSE/BSIE errors are likely to remain.⁵¹ As shown elsewhere, CCSD(T) corrections are not necessary for the backbone computations.¹⁸ The extrapolated Hartree–Fock (HF) and the MP2 correlation contributions are evaluated as follows:

$$E_X^{\text{HF}} = E_{\text{CBS}}^{\text{HF}} + A \exp(-\lambda X) E_X^{\text{Corr}} = E_{\text{CBS}}^{\text{Corr}} + BX^{-3}$$

where X is the cardinal angular momentum quantum number of the respective basis set ($X = 2$ for aug-cc-pVDZ, $X = 3$ for aug-cc-pVTZ, etc.) and parameter $\lambda = 1.43$, the value of which is optimized for the D→T variant of extrapolation. The system-unique coefficients A and B along with $E_{\text{CBS}}^{\text{Corr}}$ and E_X^{HF} , the correlation and the HF components of the total electronic energy extrapolated to CBS, respectively, need to be determined via solving listed equations, linear in all unknowns. E_X^{Corr} and E_X^{HF} terms representing the same components obtained using the aug-cc-pVXZ set of basis functions need to be inserted in the given equations.

The geometries of the methyl groups, which were added to the N1/N9 atoms of nucleobases after their detachment from the TPSS-D/LP optimized dinucleotides, were relaxed using the M06 functional along with the 6-31+G(d,p) basis set. The interaction energies of the base pairs (see Figure 4, base-to-base model) were calculated at the MP2 level of theory with a sufficiently large aug-cc-pVDZ basis set and the density fitting approximation (DF).⁶¹ We did not employ CCSD(T) to include higher order correlation effects, as they are rather insignificant for H-bonded base pairs.^{62,63} The BSSE-corrected interaction energy of a base pair ($\Delta E^{\text{M}\cdots\text{N}}$) between two interacting nucleobases M and N is defined as

$$\Delta E_{\text{BSSE}}^{\text{M}\cdots\text{N}} = E^{\text{MN}} - (E_{\text{BSSE}}^{\text{M}} + E_{\text{BSSE}}^{\text{N}})$$

where E^{MN} stands for the electronic energy of the supersystem and $E_{\text{BSSE}}^{\text{M}}$ and $E_{\text{BSSE}}^{\text{N}}$ are the BSSE-free electronic energies of the isolated subsystems obtained using the standard counterpoise procedure.⁶⁴ Since we expect the monomer deformation contributions to the interaction energies to be uniform throughout the base-pair set and since we are interested in relative energies rather than absolute values, the deformation of the monomers was neglected. The interaction energy calculations were carried out with the Molpro 2006.1 package.⁶⁵

Wave functions of the five idealized rSPSOMs (Table 1) were investigated with an atoms-in-molecules (AIM) analysis^{66–68} to reveal and compare the stabilizing effect of the conformationally specific 2'-hydroxyl H bond. Wave functions of selected base pairs were also subjected to AIM analysis with the intent to compare the strength of the interbase H bonds. AIM analyzes the local electron density curvatures and finds critical points (CPs), which can provide information on the intramolecular H-bond

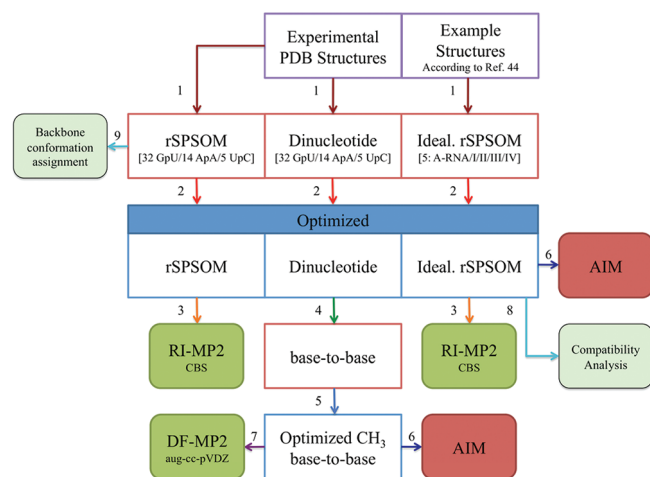


Figure 6. A symbolic flowchart describing the complete sequence of computations. The numbered steps, denoted by arrows, are as follows: (1) extraction of model systems (numerals in parentheses in the second row of boxes denote the number of systems); (2) M06/6-31+G(d,p) and subsequent TPSS-D/LP constrained geometrical optimizations; (3) calculation of RI-MP2/CBS single-point energies; (4) removal of the sugar–phosphate backbone segment followed by attachment of terminal methyl groups to N1/N9 of pyrimidines/purines; (5) relaxation of methyl groups at the M06/6-31+G(d,p) level of theory; (6) location of H-bonds via AIM analysis; (7) evaluation of DF-MP2/aug-cc-pVDZ base...base interaction energies; (8) manual attachment of nucleobases to optimized idealized rSPSOMs and inspection of mutual compatibility to form a dinucleotide platform submotif; and (9) assignment of backbone conformational class. The experimental structures out of which idealized rSPSOMs were derived are listed in the footnote of Table 1.

network. Note that the main motive for why we utilized the AIM analysis was only to reveal CPs, which give evidence of the intra- (rSPSOMs) and intermolecular (base...base) interactions. Thus, the existence of a CP between X–H and Y atoms ($X/Y = O, N, C$) gave us proof of the $X-H \cdots Y$ H bond, while its local characteristics, the electron density (ρ) and its Laplacian ($\nabla^2\rho$), measure nontrivially the strength of a given interaction. The topologies of the charge densities were computed using the converged TPSS/LP wave functions (note that the dispersion correction does not affect the wave function). The Cartesian 6d and 10f basis functions were substituted for the standard 5d and 7f functions, as recommended for the AIMPAC code.^{69,70}

The origin of initial structures as well as a summarized sequence of computations is symbolically depicted in Figure 6.

NOTATION

The individual structures are labeled as XY-z-N, where XY represents the adjacent nucleobases of the given platform (GU, AA, or UC) in the 5' → 3' direction, z stands for the respective PDB accession code, and N denotes the assigned structure number (for example, the 10 different GpU platforms identified in the large ribosomal subunit of *Haloarcula morismortui*, PDB ID 1jj2, are labeled as GU-1jj2-1 up to GU-1jj2-10). Generally, the platforms and the corresponding rSPSOM fragments are referred to as XpY, whereas the base-to-base model systems are simply labeled as XY (Figure 4). The complete list of structures considered in the current work, together with their backbone torsion

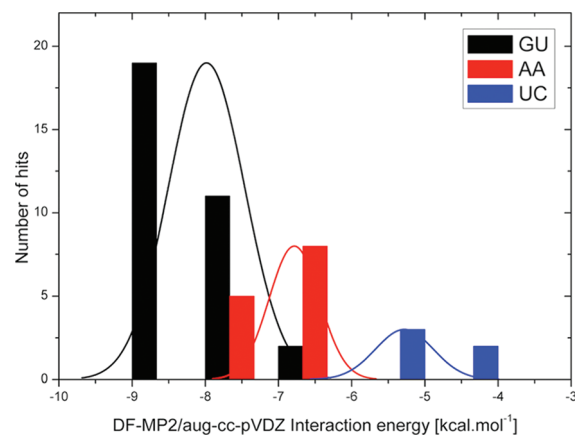


Figure 7. Distribution of the DF-MP2/aug-cc-pVDZ interaction energies of the GU (black), AA (red), and UC (blue) base pairs. The interaction energy axis is partitioned into seven bins of 1.0 kcal mol⁻¹ width. The solid curves, shown in the same colors, are the corresponding standard Gaussian normal distributions computed from the interaction energy data and thus with an infinitesimal bin width. Note that the AA interaction energies would be further reduced (in absolute value) by ~1 kcal mol⁻¹ if the artificial C9_{Met}(A)⋯N7(A) interactions were excluded. Since most of the experimental structures were determined at low atomic resolution, the energy variability primarily reflects the error of the experiments.

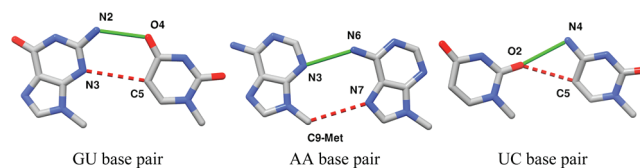


Figure 8. Sample structures of GU, AA, and UC intrastrand base pairs interacting via H bonds. The green solid lines depict the main stabilizing interactions. The weak C–H⋯O/N contacts revealed by AIM analysis are indicated by red dashed lines. Note that the C9_{Met}(A)⋯N7(A) interaction of the AA base pair is artificial (see the text). The hydrogen atoms and identified critical points (CP) have been omitted for the sake of simplicity. H bonds are depicted in a standard way, i.e., using a solid or dotted line linking the respective heavy atoms (C/O/N) without regard to the precise location of the particular CP. Note that the given CP does not need to match the geometrical line connecting the heavy atoms. More precisely, it does not even need to lie on the line connecting the H and acceptor atom nuclei. For complete molecular graphs of GU/AA/UC base pairs with all critical points, see the Supporting Information.

angles, is available in the Supporting Information, Tables S1 and S2.

RESULTS AND DISCUSSION

Base-to-Base Contribution. The base-to-base model systems (Figure 4, right) were constructed as described under Model Systems. The bases associate via a cis sugar/Hoogsteen edge interaction pattern,^{1,2} in which the 5'-nucleoside exposes its sugar edge whereas the 3'-residue exploits the Hoogsteen edge.

The interaction energies of the platform-derived GU base pairs range between -6.1 kcal mol⁻¹ (GU-1u8d-1) and -8.9 kcal mol⁻¹ (GU-2qus-1), with a mean value/standard deviation of -8.0 ± 0.5 kcal mol⁻¹ (Figure 7; histogram shown in black). The stability of the H-bonded GU base pairs is ensured by a

Table 2. H Bonds Detected by AIM Analysis to Stabilize Platform Base Pairs^a

base pair	acceptor	donor	distance	angle	density (ρ)	density Laplacian ($\nabla^2\rho$)
GU-2qus-1	O4(U)	N2(G)	2.9	164.1	0.033	0.024
	N3(G)	C5(U)	3.6	160.5	0.009	0.007
AA-1hr2-3	N3(A)	N6(A)	3.0	150.8	0.022	0.017
	N7(A)	C ⁹ _{Met} (A)	3.5	150.1	0.009	0.007
UC-1drz-1	O2(U)	N4(C)	3.0	157.8	0.019	0.020
	O2(U)	C5(C)	3.3	129.4	0.009	0.009

^aThe entries correspond to the most stable example of each type of base pair (in terms of interaction energy). The distances between heavy atoms are given in Å. The hydrogen-bonding A–H···B (A/B = O, N, C) angles are given in degrees and the charge density along with its second derivatives in a.u. The major stabilizing H bonds are highlighted in boldface. Note that the spurious C⁹_{Met}(A)···N7(A) interaction of the AA-1hr2-3 pair is not relevant to real RNA systems.

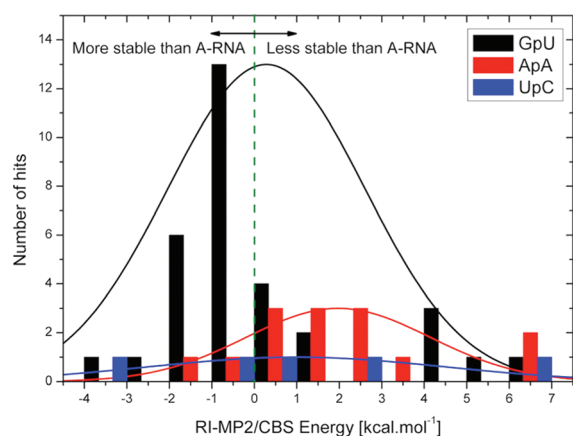


Figure 9. Distribution of the RI-MP2/CBS relative energies of rSPSOM models of GpU (black), ApA (red), and UpC (blue) experimental platform structures (i.e., idealized rSPSOMs are not included). The relative energy, reported along the horizontal abscissa, is partitioned into 12 bins of 1.0 kcal mol⁻¹ width. The curves correspond to standard Gaussian normal distributions based on the mean values and standard deviations, with infinitesimal bin width. The green dashed line at 0.0 kcal mol⁻¹ denotes the energy of the A-RNA reference state.

strong N2(G)···O4(U) H bond, which is more stabilizing than the equivalent N3(A)···N6(A) and O2(U)···N4(C) H bonds of the respective AA and UC base pairs (see the AIM results below). In addition, a second rather marginal H bond forms between N3(G) and C5(U) (Figure 8). Nevertheless, the N3(G)···C5(U) interaction is not an artifact of the selected model and might, at least slightly, contribute to base-pair stabilization.

Two intermolecular critical points, each giving evidence of a H bond, were detected in the AIM analysis of both the representative and the most stable (GU-2qus-1) GU base pair. The electron density and its Laplacian values of 0.033 and 0.024 au, respectively, at the critical point of the N2(G)···O4(U) H bond reveal this particular contact to be the strongest and the most stabilizing of the studied platform H bonds (Table 2). The second critical point between the N3(G)···C5(U) atoms, characterized by ρ and $\nabla^2\rho$ values of 0.009 and 0.007 au, indicates a weak C–H···N interaction. On the basis of our previous experience with interactions like this at the boundary between H bonds and van der Waals contacts [18], we estimate the contribution to the interaction energy to be ~ 1.0 kcal mol⁻¹ (see also the following paragraphs). Note that the AIM analysis does not allow quantitative assessment of the stabilization contributions of the respective interactions, although values of ρ should correlate with the

strength of interaction. The positive sign of $\nabla^2\rho$, which indicates depletion of the electron density at the given stationary point, signifies the ionic nature (as opposed to covalent character) of both contacts. The results are in accord with the AIM analysis of the canonical AU base pair⁷¹ in which the ratio of ρ and $\nabla^2\rho$ values of the N6(A)···O4(U) and C2(A)···O2(U) interactions are approximately 4.3 and 3.3, respectively. The equivalent ratios of the ρ and $\nabla^2\rho$ for N2(G)···O4(U) and for the weak N3(G)···C5(U) interactions stabilizing the GU-2qus-1 system are ~ 3.7 and ~ 3.4 .

The interaction energies of the AA base pairs span the range between -6.3 kcal mol⁻¹ (AA-2r8s-1) and -7.3 kcal mol⁻¹ (AA-1hr2-3) with an average value/standard deviation of -6.8 ± 0.3 kcal mol⁻¹ (Figure 7; red histogram). The AA-1hr2-5 system (-2.6 kcal mol⁻¹) is the only outlier and is accordingly excluded from the statistics. See the Supporting Information for a detailed description of this system.

The interaction energies of the AA base pairs are probably slightly biased by a weak artificial C⁹_{Met}(A)···N7(A) contact (Figure 8) that stems from N9-methylation. In order to estimate the contribution of this spurious contact to the interaction energies, we replaced the 9-methyl groups in the most stable AA-1hr2-3 base pair with hydrogens. Interaction energy calculations suggest that the unnatural contact lowers the real interaction energy by approximately 1 kcal mol⁻¹. Consequently the listed AA interaction energies are systematically shifted to lower (more stabilizing) values, and thus the actual stabilization of AA base pairs is overestimated by about ~ 1 kcal mol⁻¹.

The AIM analysis of the most stable AA-1hr2-3 system revealed two critical points, which are depicted in Figure 8 as H bonds. The electron density and its Laplacian (ρ and $\nabla^2\rho$) at the critical point of the major N3(A)···N6(A) H-bond have respective values of 0.022 and 0.017 a.u., while those of the artificial C⁹_{Met}(A)···N7(A) contact have values of 0.009 and 0.007 a.u. (Table 2).

The UC H-bonded base pairs exhibit lower stability than the GU and AA systems, with the interaction energies ranging from -4.8 kcal mol⁻¹ (UC-1u0b-1) to -5.6 kcal mol⁻¹ (UC-1drz-1) and the average value/standard deviation being -5.3 ± 0.4 kcal mol⁻¹ (Figure 7; blue histogram). Significant deviation from base–base coplanarity accompanied by elongation of the O2(U)···N4(C) distance reduces the absolute values of the energies of UC-1u0b-1 (O2···N4–3.4 Å) and UC-1vc7-1 (O2···N4–3.3 Å) interactions to -4.8 kcal mol⁻¹ and -4.9 kcal mol⁻¹, respectively. The AIM analysis of the most stable UC-1drz-1 base pair detects two critical points, one involving the O2(U)···N4(C) pair with ρ and $\nabla^2\rho$ values of 0.019 and

Table 3. RI-MP2/CBS Relative Energies (kcal mol⁻¹) and Characteristics of the H Bonds Donated by the 2'-Hydroxyl Group in Idealized rSPSOM Models of the Four Classes of Dinucleotide Platform Structures^a

class	energy	O2'–H interaction	distance	angle	density (ρ)	density Laplacian ($\nabla^2\rho$)
I	–3.2	O2'...O5'(n + 1)	3.3	140.1	0.009	0.008
		O2'...O4'(n + 1)	3.1	144.7	0.011	0.011
II	–1.1	O2'...O2P	2.9	168.2	0.032	0.022
III	+0.2	O2'...O2P	3.3	162.1	0.013	0.010
IV	+1.4	O2'...O3'	2.7	120.4	0.026	0.025
A-RNA	0.0	O2'...O3'	2.7	121.8	0.023	0.022

^aDistances between O2' and acceptor oxygens are given in Å and O2'–H...O angles in degrees. Charge densities along with second derivatives are given in a.u.

0.020 a.u. indicative of H-bonding and a second involving the O2(U)...C5(C) contact with ρ and $\nabla^2\rho$ values of 0.009 and 0.009 a.u., suggestive of a minor interaction, i.e., ~ 1 kcal mol⁻¹.

From the perspective of base-to-base contributions, stabilization of the dinucleotide platforms decreases in the following order: GU > AA > UC. The superior stability of the GU base pairs is evident from both the interaction energies and the AIM data. The average interaction energy of the GU pairs is about 1.2 kcal mol⁻¹ lower than that of the AA systems and ~ 2.7 kcal mol⁻¹ lower than that of the UC pairs. If the spurious C9_{Met}(A)...N7(A) interaction is taken into account, the difference between the GU and AA pairs is expected to reach ~ 2 kcal mol⁻¹. See the Supporting Information (Table S3) for a complete list of GU, AA, and UC base-pair interaction energies.

Sugar–Phosphate Backbone Contribution. The rSPSOM fragments of different dinucleotide platform structures (see Figure 4, center) were used to characterize the intrinsic energy preferences of the sugar–phosphate backbone. As explained under Computational Methods, the computations were carried out with constrained dihedral angles. The canonical A-RNA backbone conformation served as the reference structure with an energy of 0.0 kcal mol⁻¹ (Figure 9, the green dashed line; see Table S4 in the Supporting Information for the complete list of rSPSOM relative energies). The individual structures, which could be assigned to a known conformational class, are denoted by one of the four categories listed in Table 1 (see Table S5 in the Supporting Information for a detailed listing). Those platform backbones that could not be assigned to any class are marked as “U” (Unknown). The quality of the conformational assignment, called the “suiteness” (*S*) in the classification software,⁴⁴ is expressed as a number within the range of [0,1], where the maximum value of 1 corresponds to a structure with torsions that perfectly match those of the conformational reference state (Table 1). The systems discussed below include a tag that denotes the assigned conformational class (if any) and the *S* value, i.e., system/class/*S* value. For example, the GU-1jj2-1 rSPSOM assigned to class I (or &a following the nomenclature of Richardson et al.) with a suiteness of 0.785 would be tagged as GU-1jj2-1/I/0.785. The intrinsic rSPSOM backbone stabilities of the four classes of platform conformations are reported in Table 3.

Before introducing the data, it is important to mention that determination of the energies of the sugar–phosphate backbones of RNA crystal structures is a difficult task. Although the positions of the bases and phosphorus atoms are usually visible in the experimental electron densities, the remaining backbone atoms are rather poorly defined. Furthermore, most X-ray structures of folded RNA molecules have been solved at relatively low

resolution. This means that the individual refined backbone geometries are unavoidably affected by rather large data and refinement errors. Some of the geometries may even be unrealistic, as a consequence of averaging various substates. Thus, rather than assessing the individual geometries, we need to rely on sufficient statistics, a common practice in RNA structural bioinformatics. Unfortunately, the calculated energies are even more sensitive to data and refinement errors than the coordinates, given that energy is a highly nonlinear function of the geometries. Thus, one of the goals in this study has been to find out how much the calculated backbone energies are affected by uncertainties in the experimental structures. We assume that our results are unambiguous for the GpU platforms but less certain for the ApA platforms and even more uncertain for the UpC platforms.

The rSPSOM models derived from 32 GpU platform structures form a rather homogeneous and well-defined group of similar conformers. Roughly two-thirds of the structures (21) are more stable than the reference A-RNA structure (Figure 9; Supporting Information, Table S4), and only five outliers are appreciably less stable. The relative energies span the range -3.6 kcal mol⁻¹ (GU-1jj2-1/I/0.79) to $+7.0$ kcal mol⁻¹ (GU-2ees-1/III/0.01), with a mean value/standard error of 0.3 ± 0.4 kcal mol⁻¹. The majority of the models (especially the low-energy conformers) fall in a single conformational class (II) with relatively high *S* values (Supporting Information, Table S5) and the characteristic O2'...O2P H bond. This out-of-plane O2'...O2P H bond is enabled by the C2'-endo puckering of the S'-sugar moiety, which is one of the distinctive features of this group of structures.

One of the stable GpU backbones (GU-1jj2-1), however, adopts a distinctly different conformation from the others, with the 2'-hydroxyl group of guanosine interacting with the uridine phosphodiester O5'(n + 1) and the sugar ring embedded O4'(n + 1). This particular backbone conformer is only marginally different from canonical A-RNA topology and is the only example of an A-like platform in the model structures (Supporting Information, Table S5). Despite the structural differences compared to the typical class II GpU platform, the GU-1jj2-1 platform is embodied in a GpUpA/GpA miniduplex and thus may represent a transitional substate between A-RNA and the class II conformation. The superior stability of the idealized representation of this conformation compared to that of the predominant GpU platforms (classes I vs II in Table 3) is a consequence of a more favorable overall backbone conformation and not the specific 2'-hydroxyl H bonds. The most stabilizing interaction of the 2'-hydroxyl group is the O2'...O2P H bond observed in the majority of GpU platforms (Table 3).

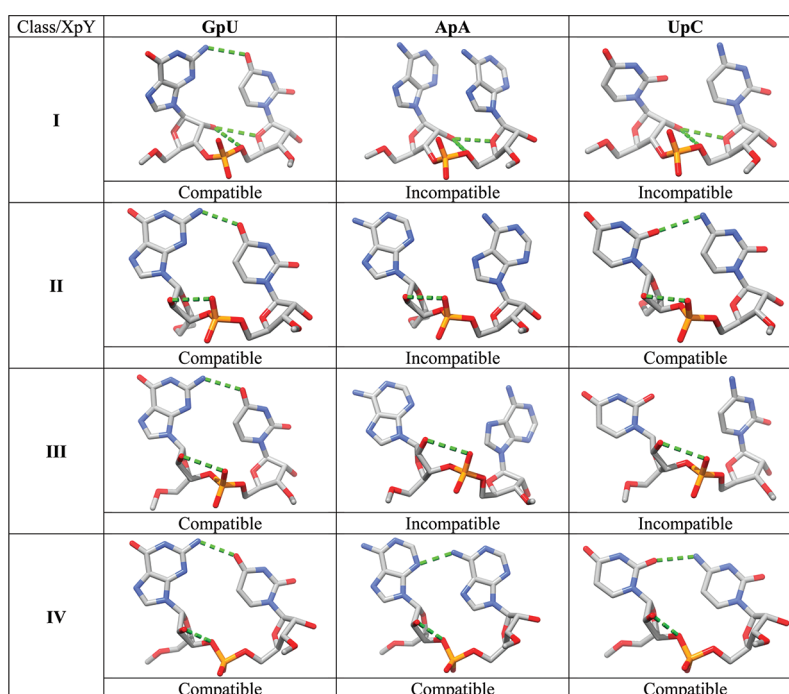


Figure 10. Atomic-level representations of 12 dinucleotide platforms created from the four identified backbone conformations (I–IV), i.e., idealized rSPSOMs, and the three studied base combinations (GU, AA, and UC). The combination is called compatible if the expected base···base interaction pattern (i.e., sugar edge/Hoogsteen edge, Figure 8) is formed. The H bonds between adjacent bases and the interactions of the 2'-hydroxyl groups typical for the conformational type (Table 3) are denoted by dashed green lines.

The five high-energy outlier systems, GU(-1q9a-1B, -483d-1B, -2ees-1, -2qus-1, and -2quw-1), bear low conformational resemblance to any of the 46 defined RNA backbone classes; i.e., the systems either could not be classified or, if assigned to a class, were characterized by a low *S* value (Supporting Information, Table S5). Three of the outliers, GU(-2ees-1, -2qus-1, and -2quw-1), are derived from medium-/low-resolution crystal structures, which likely account for the ill-defined backbone conformations. The remaining two outliers, however, are surprisingly based on high-resolution data (1q9a and 483d have a resolution of 1.0 and 1.1 Å, respectively). Both examples correspond to one of the two distinct geometries assigned to a nucleotide platform and refined with ~50% populations. The high resolution of the data seemingly provides sufficient information to attempt the refinement of two different coexisting geometries. Although the energy of one of the states is very favorable, the other is high in energy, which is a counterintuitive result. The high energy may reflect the resolution limit, i.e., representation of the data by two substates may be still insufficient to describe the backbone exhaustively. Note that the resolution does not allow consideration of more than two substates in the refinement and determination of the relative population of the two suggested substates. In addition, only a very small segment of the molecule is refined with two substates; the surrounding segments are refined assuming a single geometry. Alternatively, the structure may be correct and the energetic penalty associated with the high-energy setting of the backbone torsions may be balanced by stabilizing factors not included in the computations. See the Supporting Information for further discussion of the outliers.

The relative energies of the rSPSOM models of ApA platforms are quite diverse, with values spanning the range $-1.1 \text{ kcal mol}^{-1}$

(AA-1jj2-2/III/0.12) to $+6.2 \text{ kcal mol}^{-1}$ (AA-1hr2-3/IV/0.01) and a mean value/standard error of $2.0 \pm 0.6 \text{ kcal mol}^{-1}$ (Figure 9 and Table S4). The wide range of energies is clearly related to the fact that the backbone conformations are not uniform. Although the 3' ends of the platforms are similar, there is considerable variation in the 5' segments, particularly in the phosphodiester torsion angles and the sugar puckering. Six of the 14 observed ApA backbone geometries could not be assigned to any conformational class. The remaining eight conformers fall into two groups. The first is a slight variant of the backbone adopted by most GpU platforms and the second a conformation with highly unusual *trans* arrangements of both phosphodiester torsions (Supporting Information, Table S5). The poor match of the conformational assignments, as measured by the low *S* values, suggests that the ApA backbone substates might be either ill-defined or rarely observed. The considerable uncertainties in the experimental structures naturally affect the computed energies. The most stable ApA backbone arrangement (AA-1jj2-2) includes the strong $\text{O2}' \cdots \text{O2P}$ out-of-plane H bond, typical of GpU platforms and made possible by the same concerted changes in sugar puckering and phosphodiester linkage relative to A-RNA, i.e., C3'-exo or C2'-endo puckering of the 5'-nucleotide in combination with a *trans* ζ torsion angle. The Supporting Information includes an analysis of the three high-energy ApA outliers.

The rSPSOM models of the five UpC platforms are highly diverse. There are very few structural features common to these few examples. Only one conformer (UC-1jj2-1) can be assigned to a known conformational class, albeit with a low suitability value ($S = 0.34$). The relative energies span a wide range of values, between $-3.1 \text{ kcal mol}^{-1}$ (UC-1sj3-1) and $+6.1 \text{ kcal mol}^{-1}$ (UC-1drz-1); see Supporting Information, Table S4. It is thus

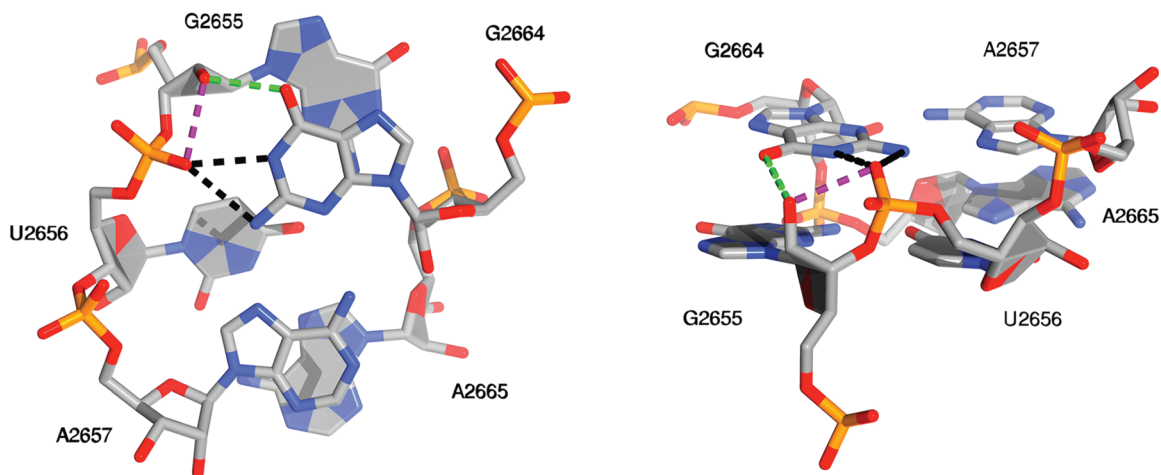


Figure 11. Atomic-level representation of the two-layered GpUpA/GpA miniduplex from the *Escherichia coli* 23S rRNA sarcin/ricin domain (PDB ID: 3dvz³⁶) showing the antiparallel 5'→3' GpUpA trinucleotide (G2655pU2656pA2657) and the nonadjacent GpA dinucleotide (G2664pA2665). The O2'...O2P (purple dashed line) forms an out-of-plane edge above the GpU platform submotif that is specifically recognized by the nonadjacent G2664 (green and black dashed lines). The nucleotides in the lower plane (the GpU platform and the nonadjacent A2665) are depicted with filled sugar and nucleobase rings, while those in the upper plane (A2657 and the nonadjacent G2664) are represented by stick models with unfilled rings. Left, view perpendicular to the GpU platform plane; right, view in the platform (lower) plane.

not possible to perform a viable statistical analysis of the energetics. Although the most energetically favored backbone arrangement (UC-1sj3-1) could not be assigned a conformational category, close examination of the torsion angles reveals its similarity to the ApA platforms with *trans* settings of both phosphodiester torsions. The advantageous setting of the backbone dihedrals cooperatively renders this conformer very stable. The fact that the given backbone conformation does not pertain to any defined conformational class, however, raises questions about its suitability within the full context of the RNA molecule. See the Supporting Information for further details. The experimental data clearly do not allow unambiguous assessment of the backbone of the UpC platform.

Compatibility of Platform Dinucleotide Sequences with a Backbone Shape. Since only some backbone conformations facilitate platform formation, we combined the three studied dinucleotide sequences (GU, AA, and UC) with the four idealized rSPSOMs that fit identified RNA conformational classes (I–IV in Tables 1 and 3). We next checked whether the given sequence is capable of forming the platform submotif. Starting with each idealized rSPSOM structure, we formed three different dinucleotide models and manually adjusted the glycosidic χ and $\chi + 1$ torsions to generate a platform-like geometry. The energy optimization was then carried out with all backbone torsions (Table 1) fixed at the initial values, i.e., at the idealized structure.

Intramolecular O2'...O2P and Interstrand Base–Phosphate H Bonds Are Cementing the Preferred Conformation of the GpU Platform in the GpUpA/GpA Miniduplex. Figure 10 shows that the 5'-GpU-3' dinucleotide can form a platform motif with all four backbone conformational classes. The A-RNA-like backbone conformation (I) enables the formation of GpU platforms with base...base interaction energies similar to those associated with the predominant mixed-pucker arrangement (II; see Supporting Information, Tables S3 and S5). The interaction energy of the bases attached to the former platform is highly favorable (~ -8.1 kcal mol⁻¹ for GU-1jj2-1). The intrinsic stability of the backbone is actually better (-3.2 kcal mol⁻¹

for the idealized class I geometry and -3.6 kcal mol⁻¹ for the GU-1jj2-1 model) than that of category II conformers (-1.1 kcal mol⁻¹, Table 3). The reason why the majority of the GpU platforms still belong to the mixed-pucker state (II) might stem from the key out-of-plane O2'...O2P H bond. The interaction not only contributes to the intrinsic stability but also forms a molecular “edge” above the base–base plane.³³ The edge is often specifically recognized by a nonadjacent guanine, as exemplified in the highly recurrent GpUpA/GpA miniduplex in the sarcin/ricin loop motif (Figure 11) where the whole GpUpA/GpA structure is stabilized by so-called 4BPh-type base–phosphate H bonding (using the nomenclature of Zirbel et al, see ref 5). The “4BPh” designation refers to a highly specific interaction between the Watson–Crick edge of a guanine and a nearby phosphate group. More specifically, the interaction in the miniduplex entails two such H bonds, N1(G)...O2P and N2(G)...O2P (black dashed lines in Figure 11), which cooperatively render the 4BPh interaction to be highly stabilizing, as well as by a sugar-phosphate H bond involving the guanines on the two strands (green dashed line in Figure 11). Therefore, it seems that the evolutionary preference for the mixed-pucker conformation of the GpU platform with its specific and well-defined backbone II arrangement may reflect a combination of good intrinsic stability and the capability to contribute to a useful and very stable RNA topology. Perhaps, the high frequency of occurrence of the GpU platform II state might also be due to hydration or water-assisted stabilization. However, presently we have no solid indications of that. The experimental structures do not suggest any unusual hydration pattern. We plan to include hydration effects into our future studies of nucleic acid backbone conformational preferences, at least in an implicit fashion.

The only backbone geometry that is compatible with formation of an ApA dinucleotide platform is the unusual conformation (IV) with *trans* arrangements of both phosphodiester torsions. This match correlates well with the conformational assignments of the experimentally determined ApA systems. Although the idealized mixed-pucker ApA backbones depicted in Figure 10 are incompatible with interbase H-bond formation,

some specific ApA dinucleotides are capable of platform formation (AA-1gid-1/3 and AA-1jj2-2 with several backbone torsions shifted away from the mean values of the conformational reference states; see Supporting Information, Table S5). Note, however, that the compatible conformation with a *trans*–*trans* phosphodiester linkage is the least stable one among the four identified conformational classes (Table 3) as a consequence of an anomeric effect, a result consistent with the higher conformational energies of all-*trans* model phosphate diesters.^{72,73} The anomeric effect is a special case of a stereoelectronic effect that disfavors *trans*–*trans* conformations of the phosphodiester linkage. The extended arrangement prevents a favorable interaction of the nonbonded electron pair on O5' with the P–O3' σ bond, and a similar interaction of O3' with the P–O5' bond. Insertion of the AA sequence into the A-like backbone substate (I) leads to a stack-like mode rather than an edge-to-edge interaction.

The UpC sequence appears to be compatible, at least on the basis of our computations, with two backbone conformational substates, the predominant mixed-pucker state (II) and the high-energy arrangement with *trans*–*trans* phosphodiester torsions (IV). The characteristic O2(U)···N4(C) contact of the UpC platform cannot be established in either the A-like conformer (I) or the alternate mixed-pucker backbone (III), as the pyrimidine bases are too far apart.

Molecular Dynamics Simulations Do Not Reproduce the Signature Interactions of the GpUpA/GpA Miniduplex and the GpU Platform. Explicit-solvent classical force field simulations represent a more common computational approach to studying RNA systems than quantum chemistry. The simplicity of the classical treatment allows for the study of rather large RNA systems with the inclusion of solvent and dynamics. The accuracy of simulations, however, is limited by the force field.⁷⁴ A reliable force field should be able to account for the highly complex backbone topology and the intricate network of molecule interactions found in the GpU platform and the GpUpA/GpA miniduplex. Indeed, the ~ 1 Å ultra-high-resolution structures of the sarcin/ricin domain containing the GpUpA/GpA miniduplex can serve as a major benchmark for force field development and testing.

A few years ago, we reported a set of what was at that time quite long, multiple 25-ns molecular dynamics (MD) simulations of the sarcin/ricin domain with the Cornell et al. AMBER force field.⁷⁵ Although the simulated system appeared basically stable, we reported some local rearrangements in the miniduplex. The backbone of the GpU platform changed in the very early stages of the simulations, with a subsequent loss of the base–phosphate H bond between the two layers of the miniduplex (the interaction denoted by the two dashed black lines in Figure 11) and a surprising shift of the glycosyl rotation of the bulged guanine from a high-anti $\chi \sim 260^\circ$ state to an even more high-anti $\chi \sim 320^\circ$ arrangement. A conformation of this sort is very unusual compared to the normal anti $\chi \sim 180$ – 200° arrangements found in RNA. The nucleotide generated in the simulations is thus subject to a large χ -dihedral internal energy penalty.^{30,32} Now, in view of our more recent experience with RNA simulations, force field tuning,^{30,32} the classification of base–phosphate interactions,^{5,6} and the present QM computations, we think that the earlier simulation results need to be reinterpreted in the following manner. The key characteristic signature interactions and the structural features of the GpUpA/GpA miniduplex were, in fact, lost in the simulations. The O2'···O2P H bond was not monitored in the earlier MD study, but it too was obviously lost.

The shift of the bulged guanine nucleotide to even higher high-anti χ values than found experimentally (formally, the simulated conformational state lies in the syn region) is evidence of a large struggle of different energy contributions in the simulated system, which evidently is not well described by the force field. Further deterioration of the system is (likely temporarily) prevented by additional interactions in the region, which lock the system close to the starting structure. We have recently performed a series of additional submicrosecond scale simulations of the sarcin/ricin domain (unpublished data), which fully confirm the above-described irreversible loss of several signature conformational features of the GpUpA/GpA miniduplex. The simulations are not improved with the latest parmbsc0³¹ and parm χ_{OL} ³⁰ variants of the Cornell et al. force field. In the near future, we will perform additional investigations of the GpUpA/GpA miniduplex with the aim to identify which force field terms may be responsible for the rearrangements seen in the simulations and to see if some tuning of the force field may be possible. The force field is presently unable to describe the highly specific and prevalent type II backbone conformation of the platform with the key out-of-plane O2'···O2P H bond.

CONCLUSIONS

A dinucleotide platform is an important noncanonical arrangement of RNA, which occurs at functionally important places in numerous molecules. The high-level quantum chemical calculations reported in this work lend credence to the hypothesis of Lu et al.³³ that the intrinsic stability of the GpU dinucleotide platform is mediated by O2'···O2P intramolecular hydrogen bonding. This conclusion is based on an analysis of the torsions and an assessment of the inherent *in vacuo* stabilities of 51 experimentally determined dinucleotide platform structures (32 5'-GpU-3', 14 ApA, and 5 UpC). We have separately studied the base···base interactions and the intervening sugar–phosphate backbone segment, each of which contributes to the overall stability.

The base···base contributions show the following stability order, GU > AA > UC. The GU pairs are, on average, ~ 2.0 kcal mol⁻¹ and ~ 2.7 kcal mol⁻¹ more stable than the AA and UC pairs, respectively. The results are supported via Bader's AIM electron topology analysis.

The GpU sugar–phosphate backbone is, on average, ~ 1.7 kcal mol⁻¹ more stable than the respective ApA conformer. We do not have enough experimental data to confidently assess the UpC platforms. Moreover, unlike ApA and UpC, the GpU backbone conformations are well-defined and fit into one of the distinct RNA conformational classes identified by Richardson et al.⁴⁴ We find the dominant GpU conformation to be more stable than the canonical A-RNA backbone and well suited to formation of a platform structure stabilized by the O2'···O2P hydrogen bond. There is, however, a rare but intrinsically even more favorable “A-like” backbone conformation, which also allows the GpU dinucleotide to take up a coplanar arrangement. In this geometry, the 2'-hydroxyl group interacts with the O5' ($n + 1$) and O4'($n + 1$) atoms in the succeeding 3'-nucleotide, rather than the anionic O2P. It, however, does not form the out-of-plane edge characteristic of most GpU platforms. The edge is very important for proper insertion of the GpU platform into the broader RNA context, as it is often recognized by a non-adjacent guanine from the opposite strand by very strong base–phosphate and base-sugar H bonds (Figure 11). The missing

edge may account for the rather infrequent incidence of “A-like” GpU platforms and the dominance of the “O2'···O2P” arrangement.

The most prevalent geometry of the GpU platform appears not to be properly described by the force fields currently used in RNA simulations and is irreversibly lost in explicit-solvent simulations.

In summary, both base···base interactions and backbone conformations enhance the stability of GpU platforms over ApA and UpC platforms. These energetic preferences correlate with the high frequency of occurrence of GpU platforms as well as with the uniform and well-defined backbone conformations. Despite the obvious limitations of our work (*in vacuo* calculations on small model systems), our successful rationalization of key features of the dinucleotide platform demonstrates that the intrinsic energy terms play important roles in determining RNA structure and sequence patterns. QM calculations thus represent a viable complement of RNA structural bioinformatics and molecular simulations in studies of the broad diversity of RNA structures.³

■ ASSOCIATED CONTENT

S Supporting Information. List of crystal structures used in the current work to derive generic GpU, ApA, and UpC dinucleotide platforms; table of the backbone and glycosidic torsion angles of the studied platforms supplemented with polar distribution plots; table of DF-MP2/aug-cc-pVDZ interaction energies of the GU, AA, and UC base pairs; table of GpU, ApA, and UpC derived rSPSOM RI-MP2/CBS relative energies; table of backbone conformational class assignment; A-RNA optimization constraints; a detailed description of the AA-1hr2-5 base pair; an in-depth analysis of GpU, ApA, and UpC derived rSPSOM outliers; and molecular graphs of GU/AA/UC base pairs. This information is available free of charge via the Internet at <http://pubs.acs.org>

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: arnost.mladek@gmail.com, sponer@ncbr.chemi.muni.cz.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENT

This work was supported by the project “CEITEC - Central European Institute of Technology” (CZ.1.05/1.1.00/02.0068) from European Regional Development Fund, by the Academy of Sciences of the Czech Republic [grant numbers AVOZ50040507 and AVOZ50040702], the Ministry of Education of the Czech Republic [grant number LC06030 and MSM0021622413], the Grant Agency of the Academy of Sciences of the Czech Republic [grant number IAA400040802], the Grant Agency of the Czech Republic [grant numbers P208/10/2302, 203/09/1476, P208/11/1822, and 203/09/H046], and the National Institutes of Health, U.S. Public Health Service [grant numbers GM096889 and GM34809]. The present study was also financially supported by the South Moravian Centre for International Mobility within the framework of the “Brno Ph.D. Talent” scholarship program, which is highly appreciated. A.M., J.E.Š., P.K., and J.Š. thank

Zdeněk Salvat for the maintenance of the computing facilities of the Brno group. The access to the MetaCentrum computing facilities provided under the research intent MSM6383917201 is also acknowledged.

■ REFERENCES

- (1) Leontis, N. B.; Westhof, E. *RNA* **2001**, *7*, 499–512.
- (2) Leontis, N. B.; Stombaugh, J.; Westhof, E. *Nucleic Acids Res.* **2002**, *30*, 3497–3531.
- (3) Sponer, J.; Sponer, J. E.; Petrov, A. I.; Leontis, N. B. *J. Phys. Chem. B* **2010**, *114*, 15723–15741.
- (4) Leontis, N. B.; Westhof, E. *Curr. Opin. Struct. Biol.* **2003**, *13*, 300–308.
- (5) Zirbel, C. L.; Sponer, J. E.; Sponer, J.; Stombaugh, J.; Leontis, N. B. *Nucleic Acids Res.* **2009**, *37*, 4898–4918.
- (6) Zgarbova, M.; Jurecka, P.; Banas, P.; Otyepka, M.; Sponer, J. E.; Leontis, N. B.; Zirbel, C. L.; Sponer, J. *J. Phys. Chem. A* **2011**, *115*, 11277–11292.
- (7) Sponer, J. E.; Leszczynski, J.; Sychrovsky, V.; Sponer, J. *J. Phys. Chem. B* **2005**, *109*, 18680–18689.
- (8) Sponer, J. E.; Spackova, N.; Kulhanek, P.; Leszczynski, J.; Sponer, J. *J. Phys. Chem. A* **2005**, *109*, 2292–2301.
- (9) Sharma, P.; Sponer, J. E.; Sharma, S.; Bhattacharyya, D.; Mitra, A. *J. Phys. Chem. B* **2010**, *114*, 3307–3320.
- (10) Mladek, A.; Sharma, P.; Mitra, A.; Bhattacharyya, D.; Sponer, J.; Sponer, J. E. *J. Phys. Chem. B* **2009**, *113*, 1743–1755.
- (11) Sponer, J. E.; Spackova, N.; Leszczynski, J.; Sponer, J. *J. Phys. Chem. B* **2005**, *109*, 11399–11410.
- (12) Sponer, J. E.; Reblova, K.; Mokdad, A.; Sychrovsky, V.; Leszczynski, J.; Sponer, J. *J. Phys. Chem. B* **2007**, *111*, 9153–9164.
- (13) Vokacova, Z.; Sponer, J.; Sponer, J. E.; Sychrovsky, V. *J. Phys. Chem. B* **2007**, *111*, 10813–10824.
- (14) Chawla, M.; Sharma, P.; Hader, S.; Bhattacharyya, D.; Mitra, A. *J. Phys. Chem. B* **2001**, *115*, 1469–1484.
- (15) Sharma, P.; Sharma, S.; Chawla, M.; Mitra, A. *J. Mol. Model.* **2009**, *15*, 633–649.
- (16) Oliva, R.; Cavallo, L.; Tramontano, A. *Nucleic Acids Res.* **2006**, *34*, 865–879.
- (17) Oliva, R.; Cavallo, L. *J. Phys. Chem. B* **2009**, *113*, 15670–15678.
- (18) Mladek, A.; Sponer, J. E.; Jurecka, P.; Banas, P.; Otyepka, M.; Svozil, D.; Sponer, J. *J. Chem. Theory Comput.* **2010**, *6*, 3817–3835.
- (19) Svozil, D.; Sponer, J. E.; Marchan, I.; Perez, A.; Cheatham, T. E.; Forti, F.; Luque, F. J.; Orozco, M.; Sponer, J. *J. Phys. Chem. B* **2008**, *112*, 8188–8197.
- (20) Mackerell, A. D. *J. Phys. Chem. B* **2009**, *113*, 3235–3244.
- (21) Foloppe, N.; Mackerell, A. D. *J. Phys. Chem. B* **1999**, *103*, 10955–10964.
- (22) Bosch, D.; Foloppe, N.; Pastor, N.; Pardo, L.; Campillo, M. *THEOCHEM* **2001**, *537*, 283–305.
- (23) Wang, F. F.; Gong, L.-D.; Zhao, D.-X. *THEOCHEM* **2009**, *909*, 49–56.
- (24) Leulliot, N.; Ghomi, M.; Scalmani, G.; Berthier, G. *J. Phys. Chem. A* **1999**, *103*, 8716–8724.
- (25) Shishkin, O. V.; Gorb, L.; Zhikol, O. A.; Leszczynski, J. *J. Biomol. Struct. Dyn.* **2004**, *21*, 537–553.
- (26) Millen, A. L.; Manderville, R. A.; Wetmore, S. D. *J. Phys. Chem. B* **2010**, *144*, 4373–4382.
- (27) Churchill, C. D. M.; Wetmore, S. D. *Phys. Chem. Chem. Phys.* **2011**, *13*, 16373–16383.
- (28) Poltev, V. I.; Anisimov, V. M.; Danilov, V. I.; Deriabina, A.; Gonzalez, E.; Jurkiewicz, A.; Les, A.; Polteva, N. *J. Biomol. Struct. Dyn.* **2008**, *25*, 563–571.
- (29) Denning, E. J.; Priyakumar, U. D.; Nilsson, L.; Mackerell, A. D. *J. Comput. Chem.* **2011**, *32*, 1929–1943.
- (30) Zgarbova, M.; Otyepka, M.; Sponer, J.; Mladek, A.; Banas, P.; Cheatham, T. E., III; Jurecka, P. *J. Chem. Theory Comput.* **2011**, *7*, 2886–2902.

- (31) Perez, A.; Marchan, I.; Svozil, D.; Spomer, J.; Cheatham, T. E., III.; Laughton, C. A.; Orozco, M. *Biophys. J.* **2007**, *92*, 3817–3829.
- (32) Banas, P.; Hollas, D.; Zgarbova, M.; Jurecka, P.; Orozco, M.; Cheatham, T. E., III.; Spomer, J.; Otyepka, M. *J. Chem. Theory Comput.* **2010**, *6*, 3836–3849.
- (33) Lu, X.-J.; Olson, W. K.; Bussemaker, H. J. *Nucleic Acids Res.* **2010**, *38*, 4868–4876.
- (34) Wimberly, B. T.; Guymon, R.; McCutcheon, J. P.; White, S. W.; Ramakrishnan, V. *Cell* **1999**, *97*, 491–502.
- (35) Correll, C. C.; Beneken, J.; Plantinga, M. J.; Lubbers, M.; Chan, Y. L. *Nucleic Acids Res.* **2003**, *31*, 6806–6818.
- (36) Olieric, V.; Rieder, U.; Lang, K.; Serganov, A.; Schulze-Briese, C.; Micura, R.; Dumas, P.; Ennifar, E. *RNA* **2009**, *15*, 707–715.
- (37) Chi, Y. I.; Martick, M.; Lares, M.; Kim, R.; Scott, W. G.; Kim, S. H. *PLoS Biol.* **2008**, *6*, 2060–2068.
- (38) Klein, D. J.; Schmeing, T. M.; Moore, P. B.; Steitz, T. A. *EMBO J.* **2001**, *20* (15), 4214–4221.
- (39) Cate, J. H.; Gooding, A. R.; Podell, E.; Zhou, K.; Golden, B. L.; Szwedczak, A. A.; Kundrot, C. E.; Cech, T. R.; Doudna, J. A. *Science* **1996**, *273*, 1678–1685.
- (40) Ke, A.; Zhou, K.; Ding, F.; Cate, J. H.; Doudna, J. A. *Nature* **2004**, *429*, 201–205.
- (41) Hauenstein, S.; Zhang, C. M.; Hou, Y. M.; Perona, J. J. *Nat. Struct. Mol. Biol.* **2004**, *11*, 1134–1141.
- (42) Lu, X.-J.; Olson, W. K. *Nucleic Acids Res.* **2003**, *31*, 5108–5121.
- (43) Lu, X.-J.; Olson, W. K. *Nat. Protoc.* **2008**, *3*, 1213–1227.
- (44) Richardson, J. S.; Schneider, B.; Murray, L. W.; Kapral, G. J.; Immormino, R. M.; Headd, J. J.; Richardson, D. C.; Ham, D.; Hershkovits, E.; Williams, L. D.; Keating, K. S.; Pyle, A. M.; Micallef, D.; Westbrook, J.; Berman, H. M. *RNA* **2008**, *14*, 465–481.
- (45) Schneider, B.; Moravek, Z.; Berman, H. M. *Nucleic Acids Res.* **2004**, *32*, 1666–1677.
- (46) Batey, R. T.; Sagar, M. B.; Doudna, J. A. *J. Mol. Biol.* **2001**, *307*, 229–246.
- (47) Klein, D. J.; Moore, P. B.; Steitz, T. A. *J. Mol. Biol.* **2004**, *340*, 141–177.
- (48) Juneau, K.; Podell, E.; Harrington, D. J.; Cech, T. R. *Structure* **2001**, *9*, 221–231.
- (49) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215–241.
- (50) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401–146405.
- (51) Jurecka, P.; Cerny, J.; Hobza, P.; Salahub, D. R. *J. Comput. Chem.* **2007**, *28*, 555–569.
- (52) Ahlrichs, R.; Bar, M.; Haser, M.; Horn, H.; Kolmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165–169.
- (53) Kendall, R. A.; Fruchtl, H. A. *Theor. Chim. Acta* **1997**, *97*, 158–163.
- (54) Feyereisen, M. W.; Fitzgerald, G.; Komornicki, A. *Chem. Phys. Lett.* **1993**, *208*, 359–363.
- (55) Vahtras, O.; Almlöf, J.; Feyereisen, M. W. *Chem. Phys. Lett.* **1993**, *213*, 514–518.
- (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 E.01; Gaussian, Inc.: Wallingford, CT, 2004.
- (57) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007–1023.
- (58) Dunning, T. H., Jr. *J. Phys. Chem. A* **2000**, *104*, 9062–9080.
- (59) Halkier, A.; Helgaker, T.; Jorgensen, P.; Klopper, W.; Olsen, J. *Chem. Phys. Lett.* **1999**, *302*, 437–446.
- (60) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *J. Chem. Phys.* **1997**, *106*, 9639–9646.
- (61) Werner, H. J.; Manby, F. R.; Knowles, P. J. *J. Chem. Phys.* **2003**, *118*, 8149–8160.
- (62) Spomer, J.; Leszczynski, J.; Hobza, P. *Biopolymers* **2001**, *61*, 3–31.
- (63) Jurecka, P.; Nachtigall, P.; Hobza, P. *Phys. Chem. Chem. Phys.* **2001**, *3*, 4578–4582.
- (64) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553–566.
- (65) MOLPRO, version 2006.1; Cardiff University: Cardiff, U. K., 2006.
- (66) Bader, R. F. W. *Atoms in Molecules. A Quantum Theory*; Oxford University Press: Oxford, U. K., 1990.
- (67) Bader, R. F. W. *Chem. Rev.* **1991**, *91*, 893–928.
- (68) Bader, R. F. W. *J. Phys. Chem. A* **1999**, *103*, 304–314.
- (69) Biegler-König, F.; Schonbohm, J.; Bayles, D. *J. Comput. Chem.* **2001**, *22*, 545–559.
- (70) Biegler-König, F.; Schonbohm, J. *J. Comput. Chem.* **2002**, *23*, 1489–1494.
- (71) Hobza, P.; Spomer, J.; Cubero, E.; Orozco, M.; Luque, J. F. *J. Phys. Chem. B* **2000**, *104*, 6286–6292.
- (72) Newton, M. D. *J. Am. Chem. Soc.* **1973**, *95*, 256–258.
- (73) Govil, G. *Biopolymers* **1976**, *15*, 2303–2307.
- (74) Ditzler, M. A.; Otyepka, M.; Spomer, J.; Walter, N. G. *Acc. Chem. Res.* **2010**, *43*, 40–47.
- (75) Spackova, N.; Spomer, J. *Nucleic Acids Res.* **2006**, *34*, 697–708.

Optimization of the CHARMM Additive Force Field for DNA: Improved Treatment of the BI/BII Conformational Equilibrium

Katarina Hart,[†] Nicolas Foloppe,[‡] Christopher M. Baker,[§] Elizabeth J. Denning,[§] Lennart Nilsson,^{†,*} and Alexander D. MacKerell, Jr.^{*,§}

[†]Department of Biosciences and Nutrition, Center for Biosciences, Karolinska Institutet, SE-141 83 HUDDINGE, Sweden

[‡]51 Natal Road, Cambridge CB1 3NY, United Kingdom

[§]Department of Pharmaceutical Sciences, School of Pharmacy, University of Maryland, 20 Penn Street, Baltimore, Maryland 21201, United States

S Supporting Information

ABSTRACT: The B-form of DNA can populate two different backbone conformations: BI and BII, defined by the difference between the torsion angles ϵ and ζ ($BI = \epsilon - \zeta < 0$ and $BII = \epsilon - \zeta > 0$). BI is the most populated state, but the population of the BII state, which is sequence dependent, is significant, and accumulating evidence shows that BII affects the overall structure of DNA and thus influences protein–DNA recognition. This work presents a reparametrization of the CHARMM27 additive nucleic acid force field to increase the sampling of the BII form in MD simulations of DNA. In addition, minor modifications of sugar puckering were introduced to facilitate sampling of the A form of DNA under the appropriate environmental conditions. Parameter optimization was guided by quantum mechanical data on model compounds, followed by calculations on several DNA duplexes in the condensed phase. The selected optimized parameters were then validated against a number of DNA duplexes, with the most extensive tests performed on the *EcoRI* dodecamer, including comparative calculations using the Amber Parm99bsc0 force field. The new CHARMM model better reproduces experimentally observed sampling of the BII conformation, including sampling as a function of sequence. In addition, the model reproduces the A form of the 1ZF1 duplex in 75% ethanol and yields a stable Z-DNA conformation of duplex (GTACGTAC) in its crystal environment. The resulting model, in combination with a recent reoptimization of the CHARMM27 force field for RNA, will be referred to as CHARMM36.

INTRODUCTION

Empirical force field based computational studies of DNA and DNA–protein complexes are of ever greater value to understand the relationship of structure and dynamics to function in these biologically essential molecules.^{1,2} This growing role for force-field-based investigations reflects increases in computational power and improvements in molecular dynamics (MD) simulation programs, allowing for longer and more relevant MD simulations of large DNA-containing systems. In this context, improvements in the force fields (FF) used to calculate the energies and forces acting on DNA are critical.³ Several different additive all-atom force fields are available for DNA, including CHARMM27,⁴ AMBER,^{5,6} Bristol-Myers Squibb,⁷ and GROMOS,⁸ where AMBER and CHARMM are the most commonly used in studies involving DNA.

While these force fields have acted as the basis for a range of successful investigations of DNA,^{9,10} and its components,¹¹ limitations in the models have surfaced. These limitations have become evident due to the ability to perform longer MD simulations on a wider variety of DNA and new experimental data that can be used to test force fields.^{12,13} For example, a problem was identified in simulations >10 ns involving treatment of the α and γ dihedrals in the phosphodiester backbone with AMBER.^{14,15} This problem was solved on the basis of quantum mechanical (QM) calculations on model compounds representative of the phosphodiester backbone used to direct parameter optimization,

followed by extensive MD simulations. Other improvements in the AMBER force field important for DNA simulations have involved the ions,¹⁶ and various adjustments in the AMBER χ parameters have also been presented.^{17–19}

With the CHARMM DNA force field, limitations in the treatment of the relative populations of the BI and BII substates of the canonical B form of DNA have been noted, where the BII state is significantly underestimated relative to the BI state.^{13,20,21} The BI and BII states are defined on the basis of the phosphodiester torsions ϵ and ζ . BI is characterized by $\epsilon - \zeta$ around -90° and BII by $\epsilon - \zeta > 0$. The BII conformation was first characterized in a crystal structure²² and subsequently observed using ³¹P NMR chemical shifts and scalar coupling constants.^{20,23,24} NMR data led to quantification of the intrinsic sequence-specific propensities to populate BII in solution for every DNA base step.²⁵ Crucially, the BI/BII equilibrium affects the DNA helical parameters, especially the twist, roll, and base-pair displacement from the main helical axis. This influences the DNA overall structure. For instance, it explains to a large extent the sequence-specific variations in B-DNA groove dimensions.²⁶ Approximately 20% of the base steps in free DNA significantly populate BII and somewhat less in protein bound DNA.²⁷ The BI/BII equilibrium has consequences for DNA recognition by

Received: October 13, 2011

Published: December 07, 2011

proteins involved in sequence specific binding^{21,28} as well as when the binding is nonspecific or uses an indirect readout mechanism.²⁵ The sequence-dependent propensity to adopt the BII state has been suggested to contribute to the ability of DNA sequences to form nucleosomes.²⁵ Changes upon going from the BI to BII state alter the solvent accessibility of backbone atoms; for instance, in BI the O3' is accessible, but not in the BII conformation. Furthermore, the BI/BII equilibrium is also sensitive to the composition (Na^+ or K^+) of the ionic environment at physiological concentrations,²⁹ which might provide another mechanism to tune protein–DNA recognition. The energy barrier between the BI and BII states is of interest, although the current estimates strongly depend on the model used.^{24,30,31}

Underestimation of the BII state by the CHARMM27 FF has been observed in a number of studies. MD simulations of transcription factor Ndt80 in complex with DNA²¹ showed the FF to not reproduce details of the crystallographic conformation of the DNA, in particular the BII state around the crucial base step T6'–G5'.^{28,32} Notably, the inability to sample the BII conformation led to the simulated protein side chains forming different interactions with the DNA, compared to the X-ray structure. Another system is the JunFos DNA oligomer, which has distinct BII populations.²⁰ JunFos has been used to develop an NMR-based method to quantify the BII populations at every phosphodiester linkage.²⁰ Subsequently, MD simulations with CHARMM27 and the Parm99bsc0¹⁵ and Parm98⁶ versions of the AMBER FF showed that none of the FFs could reproduce the experimental BI/BII populations in the absence of NMR restraints.¹³ These observations, and the fact that sampling of BII states is significant in a range of sequences,^{25,33} motivated the present fine-tuning of the CHARMM27 DNA FF. As presented below, this involved the systematic optimization of the dihedral parameters associated with the ϵ and ζ torsions as well as the C2'–C3'–C4'–O4' torsion that influences the relative energies of the north and south sugar ring puckers. The resulting modified CHARMM DNA force field yields significant improvement in the treatment of the BI/BII equilibrium. In addition, the modified force field has improved sensitivity of the DNA to its environment, better reproducing the change from the A form of DNA to the B form as a function of water activity due to the presence of ethanol. The FF is also shown to satisfactorily reproduce the structure of a Z-form duplex in its crystal environment. The modified parameters, which are not applicable to RNA, will be included in the new CHARMM36 force field for oligonucleotides, alongside a recent enhancement in the treatment of RNA.³⁴

METHODS

Quantum Mechanical Calculations. QM calculations were performed with the programs Gaussian 09³⁵ and QChem³⁶ on the model compounds shown in Figure 1. As previously presented for model 1,³⁷ structures were optimized at MP2/6-31(+)-G(d) to default tolerances in Gaussian, and single-point energies were calculated at the RI-MP2/cc-pVTZ level with QChem. Optimizations were initiated with selected dihedral angles in the sugar and phosphate moieties, constrained to values obtained from statistical surveys of DNA crystal structures in the protein³⁸ and nucleic acid databases,³⁹ as previously described.⁴⁰ For model compound 1, following the initial constrained optimizations, additional optimization of sugar pucker with the

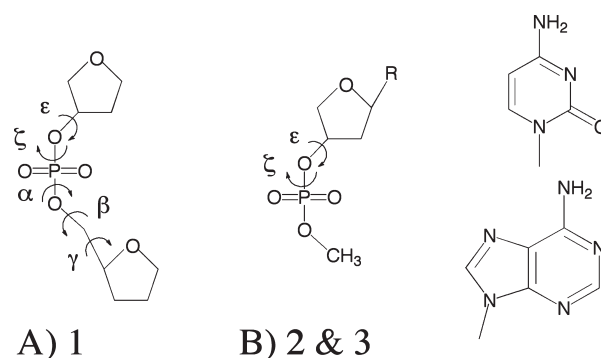


Figure 1. Model compounds used for parameter optimization. In B, the base, R, is either cytidine for model 2 or adenine for model 3, as shown.

backbone dihedral constraints maintained was performed at the MP2/6-31(+)-G(d) level followed by the RIMP2 single point calculations. Model compounds 2 and 3 were optimized with only a single sugar ring dihedral constrained, following which the value of the sugar pucker was extracted as described by Foloppe and MacKerell.⁴ For the ϵ versus ζ BI/BII 2D surface on the analog of model compound 2 lacking a base, optimizations were performed at the MP2/6-31(+)-G(d) level with the single sugar dihedral and α dihedral restraints corresponding to B form DNA, with ϵ and ζ sampled in 15° increments; no single point RIMP2 energies were obtained for this system.

Crystallographic Survey. Target data for FF validation were obtained, in part, from a survey of the crystal structures in the protein databank.³⁸ Included in the survey were only double helical DNA duplexes with unmodified DNA bases or backbones, no protein or RNA in the structure, and with a resolution ≤ 2.5 Å. Terminal nucleotides were excluded from the analysis unless noted.

Molecular Mechanical Calculations. MM calculations were performed with the programs CHARMM⁴¹ and NAMD⁴² using the CHARMM27 all-atom nucleic acid force field^{4,43} with the modifications discussed below. All systems were solvated with the TIP3P water model,⁴⁴ a minimum of 8 Å beyond the solute non-hydrogen atoms, and made electrically neutral by the addition of Na^+ ions (1xnp, *EcoRI* and JunFos) or Mg^{2+} ions (BDJ025 and GTAC2). With the *EcoRI* excess salt simulation, additional Na^+ and Cl^- ions (5 Na^+ and 5 Cl^-) yielded a concentration of 100 mM NaCl. Subsequent calculations were performed with periodic boundary conditions. Crystals were solvated and pre-equilibrated as previously described.⁴ Systems were first energy minimized for 500 adopted-basis Newton–Raphson (ANBR) steps in the presence of harmonic restraints of 5 kcal/mol/Å² on the solute non-hydrogen atoms, followed by 20 ps NPT simulations in the presence of those restraints. Then, an additional 500 ANBR step minimization in the absence of harmonic restraints was performed, after which the production MD simulations were initiated. Unless noted, MD simulations were performed for 100 ns (Table 1) at 298 K and a pressure of 1 atm, using Hoover temperature control⁴⁵ and the Langevin piston to maintain the pressure.⁴⁶ The integration time step was 2 fs, and SHAKE⁴⁷ was used to constrain X–H bonds during the simulations. For selected systems (Table 1), a lookup table was used in the evaluation of nonbonded interactions in order to speed up the simulations.⁴⁸ Electrostatic interactions were calculated using the particle mesh Ewald method (PME)⁴⁹ with a κ value of 0.36 and a real space cutoff of 10 or 12 Å (Table 1).

Table 1. DNA Systems Used for Parameter Training, Tests, and Validation^a

sequence	comment/reference
(1) d(GTACGTAC)*	GTAC, A form crystal (ADH059) ⁶⁹
(2) d(CGATCGATCG)*	BDJ025, B form crystal (BDJ025) ⁷⁰
(3) d(CGCGAATTCGCG)	<i>EcoRI</i> dodecamer., MD to 300 ns with C27_2b, X-ray/NMR ^{71–77}
(4) d(GCATTCTGAGTCAG)*	JunFos, experimental BII content ^{13,20,29}
(5) d(GAAGAGAAGC)*	1AXP, NMR, high purine content strand ⁷⁸
(6) d(ACACTACAATGTTGCAAT)	3BSE, B form, X-ray 1.60 Å; disordered region ⁶⁵
(7) d(CCGTCGACGG)	1ZF7, B form, X-ray (1.05 Å) ⁷⁹
(8) d(CCGGGCCCGG)	1ZF1, A form, X-ray (1.35 Å) ⁷⁹
(9) d(TGCGCA)	1LJX, Z form, X-ray (1.64 Å) ⁸⁰
(10) d(GACTTTCAGGG)*	NF-κB, B form, NMR, ⁶² experimental BII content
(11) d(TGCGACACAAAACT)*	Ndt80 binding site, complex with protein, X-ray (1.4 Å) ²⁸

^a All validation simulations were performed for 100 ns unless noted. Comment/Reference includes the PDB or NDB identifiers. Starting structures for the simulations were the crystal structures except with #5 1AXP where the NMR structure was used and #3 *EcoRI* and #4 JunFos, where the canonical B form was used as the starting structure. Systems indicated with * were simulated using the non-bond lookup table in CHARMM.⁴⁸

Lennard-Jones interactions were truncated at the same distance as the PME real space cutoff in the respective simulations, with smoothing over the last 2 Å using the force switch method.⁵⁰ Nonbond atom pair lists were updated heuristically whenever any atom moved more than half the distance between the list cutoff (CUTNB) and the interaction cutoff (CTOFNB) distances.

Analysis was performed on coordinates saved every 5 ps from the MD simulations unless noted. Dihedral angle distributions were analyzed using 5° bins and the following torsion definitions: $\alpha = \text{O}_{3'}-\text{P}-\text{O}_{5'}-\text{C}_{5'}$, $\beta = \text{P}-\text{O}_{5'}-\text{C}_{5'}-\text{C}_{4'}$, $\gamma = \text{O}_{5'}-\text{C}_{5'}-\text{C}_{4'}-\text{C}_{3'}$, $\varepsilon = \text{C}_{4'}-\text{C}_{3'}-\text{O}_{3'}-\text{P}$, and $\zeta = \text{C}_{3'}-\text{O}_{3'}-\text{P}-\text{O}_{5'}$. Helicoidal analysis used the Curves^{51,52} package with data placed in 5° or 0.2 Å bins. Root-mean-square (RMS) differences were calculated with respect to the canonical (A, B, or Z) DNA forms of the respective sequences, unless noted, following alignment of all non-hydrogen atoms. Unless noted, the terminal nucleotides were excluded from the analyses. BI versus BII populations from the MD simulations were obtained by simple counting, i.e., BI if $\varepsilon-\zeta < 0$ and BII if $\varepsilon-\zeta > 0$; however, this method differs from that used to obtain BII population estimates from NMR, as discussed below.

RESULTS AND DISCUSSION

In the present work, a systematic optimization of the CHARMM27 all-atom additive force field for nucleic acids^{4,43} was undertaken to improve the ability of the model to represent the relative populations of the BI and BII conformers of DNA. To achieve this without significantly altering the remainder of the FF (e.g., the treatments of Watson–Crick base pair interactions, which have been shown to yield good agreement with

NMR experiments with respect to base flipping^{53–55}), the optimization focused on the dihedral parameters associated with the ε and ζ torsions in the phosphodiester backbone. In addition, it was necessary to modify the relative energies of the north and south puckers of the deoxyribose sugar, targeting the C2'–C3'–C4'–O4' associated dihedral parameter, to allow for sampling of A-form DNA in the appropriate conditions. In the following, results for model compounds on which QM data are available are presented for the CHARMM27 FF (C27) and for five modified parameters sets, which were used in preliminary simulations of three systems (Table 1, systems 1–3). The final selected set of parameters was then used to simulate additional systems (Table 1, systems 4–11) to more rigorously test the force field. Altogether, 3.4 μs of simulation was performed on 11 DNA duplexes of different compositions and sizes (total system sizes range from 1200 to 51 000 atoms) with explicit solvent, in crystal or solution, and also bound to a protein.²¹

Model Compound Calculations. Model compounds representative of the backbone and nucleotide unit in DNA on which QM data are available are shown in Figure 1. The first model, **1**, which contains two furanoses connected by a phosphodiester linkage, has been subjected to extensive QM calculations.³⁷ The conformational energies of compound **1** as a function of its dihedrals reflected the dihedral distributions from survey data on crystal structures of duplex DNA, thereby validating it as a model for the phosphodiester backbone. Accordingly, model compound **1** was used for optimization of the ε and ζ dihedral parameters. The other model compounds, **2** and **3**, are cytosine and adenosine nucleosides, respectively, used previously in the optimization of the C27 nucleic acid FF. As the goal of the present study was to perform minimal adjustments of the C27 FF, use of two nucleosides to evaluate changes in sugar pucker energies was deemed sufficient. As shown below, the changes to the sugar parameters only had a minor impact on the relative energies of the north (C3'endo) vs south (C2'endo) conformations, while the overall energy as a function of pucker was not significantly changed.

Table 2 presents the relative energies of the BII conformation of **1** with respect to the BI conformation. As may be seen, C27 significantly overestimates the QM result. Accordingly, initial efforts aimed to alter only the BI/BII energy difference by altering just the ζ dihedral to yield parameter set C27_1, and then both ε and ζ in set C27_2; adjusted parameters for all sets are shown in Table S1 of the Supporting Information. Adjustments in C27_1 led to improved agreement with the QM target data, but the C27_2 modification further lowers the relative energy and further improves agreement with the QM value. A third set, C27_3, was developed in which only ζ parameters were again modified, yielding a relative energy slightly lower than that of the C27_2 set. As shown below, these initial sets yielded improvements in sampling of the BII state; however, they overly destabilized the A-form of DNA.

The ε and ζ potential energy surfaces for **1** are shown in Figure 2 from QM calculations as well as from C27 and from the final, selected parameter set, C27_2b (see below). The overall shape of the empirical surfaces mimics that of the QM surfaces; however, differences in the relative energies of different minima are evident. These differences are due to limitations in the ability of the MM energy function to reproduce all of the details of the QM energy surfaces and, more importantly, systematic shifts in the C27 energy surface relative to the QM surfaces implemented to yield better sampling in oligonucleotide simulations as judged

by the reproduction of crystal survey data.⁴ Such deviations from QM energies are necessary for the FF to improve agreement with the condensed phase properties for oligonucleotides given the inherent limitations in the potential energy function as well as the challenges of parameter optimization.

To account for the inability of the first modified parameter sets to both adequately sample the BII conformation and maintain

Table 2. Relative Energies of the BI and BII Conformations of Model Compound 1 Including the Minimized Values of the Dihedral Angles^a

level of theory	ΔE	puck1	ϵ	ζ	$\alpha + 1$	$\beta + 1$	$\gamma + 1$	puck2
B _I								
QM	0.00	8.3	201.1	281.6	288.5	181.6	46.6	-4.1
C27	0.00	9.0	189.3	267.0	302.2	170.4	28.2	-2.7
C27_1	0.00	9.2	191.5	267.5	302.2	170.1	28.4	-2.7
C27_2	0.00	9.3	190.0	265.2	302.9	169.9	27.8	-3.0
C27_2b	0.00	7.6	190.2	264.3	304.0	172.4	22.7	-9.6
C27_3	0.00	8.9	188.0	266.8	302.2	170.7	28.0	-2.8
C27_3b	0.00	7.0	187.7	266.0	303.1	173.0	23.5	-9.1
B _{II}								
QM	0.97	3.0	267.1	167.4	289.6	242.5	49.0	-6.6
C27	2.78	15.7	260.5	184.2	293.1	171.0	51.0	-4.9
C27_1	1.91	16.1	262.8	184.6	292.9	170.9	51.0	-4.9
C27_2	1.59	15.8	261.7	183.6	293.0	171.0	51.0	-4.9
C27_2b	1.37	16.3	262.3	183.7	293.0	170.7	51.0	-10.7
C27_3	1.49	17.4	267.6	186.5	292.4	170.8	51.0	-4.9
C27_3b	1.49	18.3	267.8	186.7	292.4	170.5	51.0	-10.8

^aEnergies in kcal/mol and angles in degrees. QM relative energies were obtained at the MP2/6-31(+)-G(d)//RIMP2/cc-pVTZ level. Pucker is represented by the C1'-O4'-C4'-C3' dihedral angle for the first (Puck1) and second (Puck2) sugar in model compound 1.

the A-form of DNA in its crystal environment (see below), modifications of the sugar dihedral parameters were undertaken. These modifications were based on the adenine and cytosine nucleosides (Figure 1b) and only involved a single dihedral (C2'-C3'-C4'-O4') in the furanose ring (Table S1, Supporting Information). Shown in Table 3 are the relative energies of the north pucker relative to those of the south pucker for the two nucleosides. The C27 puckers were empirically adjusted to reproduce the north versus south distributions from a crystal survey of the nucleic acid database, with QM MP2/6-31G(d) relative energies also used as a guideline. The resulting empirical model overestimates the QM values of the north pucker for both model compounds; the lower energy of the north pucker of the cytosine nucleoside is consistent with cytidine bases favoring the A form of DNA, as previously discussed.⁵⁶ Once it was observed that the ϵ and ζ dihedral parameter modifications led to destabilization of the A form of DNA (see below), the north pucker energy was lowered relative to that of the south by 0.7–0.8 kcal/mol. The resulting model of the sugar now underestimates the QM relative north pucker energy by 0.3 and 0.1 kcal/mol for the adenine and cytosine nucleosides, respectively (Table 3). The altered sugar energetics were combined with the C27_2 set to give the C27_2b set, which yielded satisfactory simulated sugar pucker distributions for duplex DNA in condensed phase, as shown below. The same sugar parameter modification was also applied to C27_3, yielding C27_3b.

Initial Parameter Set Selection Based on MD Simulations.

Selection from the parameter sets developed in the preceding section was based on condensed phase simulations of three DNA duplexes, GTAC2, BDJ025, and the *EcoRI* dodecamer (Table 1)). GTAC2 and BDJ025, which are in the A and B forms, respectively, were simulated in their crystal environments, thereby allowing for a more rigorous comparison of the experimental and simulation data. Special care was taken to maintain the A form of GTAC2. Inclusion of the *EcoRI* dodecamer in solution was based on the central role that the duplex has played as a benchmark in computational studies of DNA^{4,57–59}

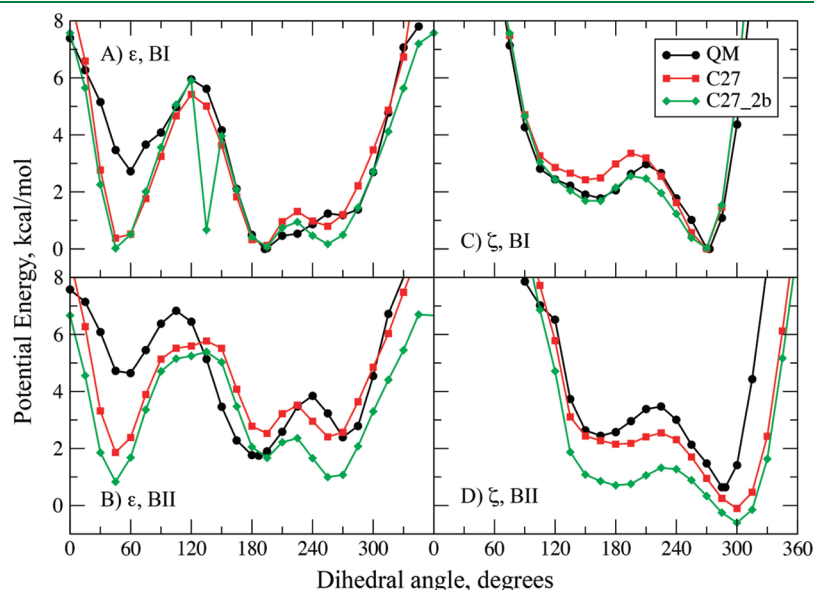


Figure 2. Potential energies as a function of the ϵ (A and B) and ζ (C and D) dihedrals for model compound 1 with the remainder of the rotatable bonds maintained at dihedral angles corresponding to the BI (A and C) or BII (B and D) canonical conformations. Note that in the surface in panel A with C27_2b, a local minimum was present at 135°; this minimum in the potential energy surface was not observed to impact MD simulations of the DNA duplexes.

Table 3. Relative Energies of the North and South Conformations of the Nucleosides of Adenine and Cytosine (Model Compound 2)^a

level of theory	north		south	
	phase angle	ΔE	phase angle	ΔE
NUSA				
QM	10.7	1.03	169.2	0.00
C27	9.4	1.40	167.1	0.00
C27_1	9.4	1.40	167.1	0.00
C27_2	9.4	1.40	167.1	0.00
C27_3	9.4	1.40	167.1	0.00
C27_2b	12.3	0.69	168.7	0.00
C27_3b	12.3	0.69	168.7	0.00
NUSC				
QM	10.3	-0.49	164.0	0.00
C27	9.6	0.18	167.1	0.00
C27_1	9.6	0.18	167.1	0.00
C27_2	9.6	0.18	167.1	0.00
C27_3	9.6	0.18	167.1	0.00
C27_2b	12.3	-0.60	168.7	0.00
C27_3b	12.3	-0.60	168.7	0.00

^aEnergies in kcal/mol and angles in degrees. Sugar pucker phase angle calculated on the basis of the method of Altona and Sundaralingam⁸¹ with QM data at the RIMP2/cc-pVTZ//MP2/6-31(+)G(d) level.

and on the need to include solution data to assess the behavior of the parameter sets with respect to Watson–Crick (WC) base pair interactions, sugar pucker, as well as the BI/BII sampling in the phosphodiester backbone in a high water activity environment.

Populations of the BI and BII states for BDJ025 and *EcoRI* averaged over all of the nucleotides in the duplexes for the different parameter sets are in Table 4. Such analysis did not include GTAC2 at this stage of the study as A-form DNA does not populate the BII conformation. As expected from previous investigations, the BI/BII equilibrium with C27 in BDJ025 and *EcoRI* is dominated by the BI state. Alteration of the parameters based on model compound 1 to lower the relative energy of the BII conformation yielded the anticipated increase in sampling of that state for both systems. The C27_2 and C27_3 sets give additional sampling of BII, consistent with the lower energy of the BII conformation in model compound 1 (Table 2). Given that experimental studies indicate the overall population of the BII state to be 37% in *EcoRI* in solution,²⁴ the sampling of BII by sets C27_2 and C27_3 indicated that they provided a reasonable basis to develop the final model.

As increasing the BII population was readily achieved, the simulations were analyzed with respect to the ability of the models to reproduce the A and B conformations of DNA. Average RMS differences versus canonical forms of DNA for all non-hydrogen atoms in non-terminal residues are presented in Table 5. The B form conformation (BDJ025 and *EcoRI* duplexes) was maintained for all of the parameter sets. With A form GTAC2 with sets C27_2 and C27_3, there is a tendency for the RMS difference versus the A form to increase as compared to that occurring with C27. Detailed analysis of these trajectories (not shown) indicated that the shift away from the A form was associated with the sugars

Table 4. Population of the BI and BII States for the *EcoRI* Dodecamer and BDJ025 over the Full Oligonucleotides for All Parameter Sets Tested^a

parameter set	<i>EcoRI</i>		BDJ025	
	BI	BII	BI	BII
C27	0.89	0.11	0.83	0.17
C27_1	0.75	0.25	0.67	0.33
C27_2	0.72	0.28	0.60	0.40
C27_2b	0.74	0.26	0.62	0.38
C27_3	0.70	0.30	0.59	0.41
C27_3b	0.79	0.21	0.65	0.35

^aResults obtained over 40 or 80 ns simulations for *EcoRI* and BDJ025, respectively, with the average over all of the ϵ, ζ pairs for *EcoRI*, and with BDJ025 the average populations were obtained for each nucleotide with the presented values being the average of those values.

switching from the typically A-form north pucker to the typically B-form south pucker. Accordingly, selected sugar dihedrals were adjusted to lower the relative energy of the north conformation, also yielding better agreement with QM data on the adenine and cytosine nucleoside model compounds (Table 2). For C27_2, an additional adjustment of the ϵ dihedral parameter was undertaken (Table S1, Supporting Information), leading to a slight lowering of the energy of the BII conformation in model compound 1.

The two parameter sets with the modified sugar parameters, C27_2b and C27_3b, were then tested in simulations of the three training set oligonucleotides. Both models maintained the A conformation of GTAC2 as well as the B conformations of *EcoRI* and BDJ025 (Table 5). Consistent with the BII conformational energies in model compound 1, C27_2b sampled the BII state more than C27_3b (Table 4), with C27_2b yielding an overall BII population closer to the 37% seen in experimental studies of *EcoRI*. Accordingly, parameter set C27_2b was tested further with simulations of additional DNA duplexes for a more rigorous and general validation.

Further Tests and Validation of the C27_2b Parameter Set. Additional testing and validation of the C27_2b parameter set was performed via extending the simulations on the *EcoRI*, BDJ025 and GTAC2 duplexes as well as performing simulations on additional systems (Table 1). The additional systems were selected to vary in sequence and size, to be of high crystallographic resolution, or to have been subjected to analysis of BII content in NMR studies. For example, the *EcoRI* (#1), JunFos (#4), and NF- κ B (#10) sequences have been subjected to explicit analysis of the BII content in solution. A longer oligonucleotide (3BSE, #6) was selected in part due to unique dynamic aspects of the molecule in solution. Sequences (1ZF7 #7 and 1ZF1, #8) were of interest, as they crystallize in the B and A forms, respectively, despite their similar sequences. 1LJX (#9) was selected as it is in the Z form; this system along with GTAC2 and BDJ025 were simulated in their explicit crystal environments while the remaining systems were simulated in solution. The oligonucleotide targeted by NF- κ B (#10) was selected due to the availability of experimental BII populations on that sequence. This system was also the first in which underestimation of the BII state in MD simulations was noted by us. The Ndt80 sequence (#11) was also of interest as its BII conformation at a specific nucleotide is important for interaction with the protein; this system was one in which the inability to properly sample the BII conformation was noted.²¹

Table 5. Average RMS Differences (Å) with Respect to the Canonical Forms of DNA for the EcoRI Dodecamer, BDJ025, and GTAC2 Oligonucleotides for All Parameter Sets Tested^a

parameter set	EcoRI		BDJ025		GTAC2	
	vs A	vs B	vs A	vs B	vs A	vs B
C27	4.26 ± 0.46	2.10 ± 0.40	3.77 ± 0.21	1.73 ± 0.17	1.81 ± 0.14	3.56 ± 0.17
C27_1	5.66 ± 0.45	2.33 ± 0.39	4.24 ± 0.24	1.53 ± 0.13	1.56 ± 0.11	3.37 ± 0.20
C27_2	5.62 ± 0.39	2.18 ± 0.32	4.78 ± 0.20	1.40 ± 0.12	2.10 ± 0.19	2.90 ± 0.25
C27_2b	4.46 ± 0.51	2.19 ± 0.40	4.21 ± 0.19	1.72 ± 0.13	1.69 ± 0.14	3.64 ± 0.15
C27_3	5.80 ± 0.47	2.29 ± 0.49	4.65 ± 0.24	1.47 ± 0.13	2.21 ± 0.12	2.40 ± 0.23
C27_3b	4.80 ± 0.52	2.16 ± 0.38	4.71 ± 0.19	1.43 ± 0.11	1.79 ± 0.14	3.54 ± 0.21

^a Results, over all non-hydrogen atoms in non-terminal residues, obtained over 20 to 40 ns for EcoRI and 60 to 80 ns for BDJ025 and GTAC2. Errors represent the RMS fluctuations

Table 6. Average RMS Differences (Å) with Respect to the Canonical Forms of DNA for the C27_2b Validation Simulations^a

system	C27		C27_2b		AMBER bsc0	
	vsA	vsB	vsA	vsB	vsA	vsB
EcoRI ¹	4.21 ± 0.46	2.14 ± 0.43	4.09 ± 0.56	2.42 ± 0.45	5.27 ± 0.49	2.37 ± 0.35
			5.05 ± 0.53	2.31 ± 0.45		
GTAC2	1.92 ± 0.19	3.73 ± 0.23	1.68 ± 0.14	3.53 ± 0.22		
BDJ025	3.92 ± 0.25	1.66 ± 0.17	4.13 ± 0.22	1.72 ± 0.14		
1AXP	3.71 ± 0.35	1.58 ± 0.31	4.02 ± 0.57	2.42 ± 0.58		
3BSE	4.26 ± 0.55	2.77 ± 0.59	5.21 ± 0.65	3.71 ± 0.91		
1ZF7	3.62 ± 0.42	1.52 ± 0.31	4.23 ± 0.57	2.07 ± 0.51		
1ZF1(H ₂ O)	3.50 ± 0.43	1.67 ± 0.34	4.06 ± 0.50	2.07 ± 0.52		
1ZF1(EtoH)	3.27 ± 0.47	1.68 ± 0.41	1.34 ± 0.22	4.43 ± 0.31		
JunFos	4.24 ± 0.40	1.58 ± 0.26	6.06 ± 0.69	2.61 ± 0.57	5.38 ± 0.56	3.20 ± 0.65
NF-κb	5.02 ± 0.58	2.34 ± 0.58	5.50 ± 0.61	2.32 ± 0.54		
Ndt80			6.47 ± 0.26	2.28 ± 0.21	6.09 ± 0.39 ²	2.40 ± 0.29 ²
1LJX(Zform) ³	1.14 ± 0.08		1.18 ± 0.10			

^a Data averaged over 20 to 100 ns except for 1AXP C27 (20–98 ns), Ndt80 C27_2b (10–50 ns), and Ndt80 Amber parm94 (1–10 ns). Errors represent the RMS fluctuations about the average. (1) The second row of C27_2b results for EcoRI are from the high salt simulation. (2) Trajectory from Hart and Nilsson²¹ where the AMBER parm94 parameters⁵ were used. (3) Results for 1LJX are with respect to the crystallographic structure.

With EcoRI and JunFos, simulations were also performed with the AMBER Parm99bsc0 force field^{5,15} to allow comparison with the final parameter set developed in the present study; this will be referred to as AMBER for the remainder of the manuscript, unless noted. In this section various quantities from simulations of these systems are presented, including comparison of the C27_2b FF results with experimental data and with the C27 FF. The C27_2b results are also compared with AMBER for EcoRI and JunFos. Some results are presented in the Supporting Information. Given the extensive amount of literature on the EcoRI dodecamer, the majority of presented data is on that system.

The first test examined the RMS differences (RMSD) between the simulated duplexes and their corresponding canonical A and B forms. In all cases, the outcomes were consistent with the experimental data (Table 6). For EcoRI, 1AXP, 3BSE, 1ZF7, JunFos, and NF-κB, the conformations in solution are closer to the B form versus the A form. With EcoRI, a second simulation at higher salt (100 mM NaCl, see below) yielded an average conformation which showed no tendency to shift toward the A form; this result is consistent with experimental results since the salt concentration is well below that required to stabilize the A form of DNA. The BDJ025 crystal simulation is also closer to the

B form, as expected. Of note are the results for the A form structures (GTAC2 and 1ZF1). The GTAC2 crystal simulation yields a structure close to the canonical A form, as does the 1ZF1 simulation in 75% ethanol for the C27_2b FF, consistent with expectations for a GC-rich duplex in a low water-activity environment. In contrast, the simulation of 1ZF1 in 75% ethanol using C27 converts to the B form (see Figure S1, Supporting Information). Importantly, the simulation using C27_2b converted to the B form in the water (0% ethanol) simulation, consistent with the impact of high water activity on DNA conformation.⁶⁰ These results indicate that C27_2b is more sensitive than C27 to changes in the water activity environment, a critical and stringent test for DNA force fields. Additional testing of this phenomenon is warranted in future studies.

To robustly test the stability of C27_2b, the EcoRI simulation was extended to 300 ns (Figure 3) during which the duplex remained close to the B form. The Watson–Crick base pairing as defined by the N1···N3 distance is well maintained in the simulation for all of the FFs (inset in Figure 3). Interestingly, the N1···N3 distance distributions from the simulations peak at a slightly longer distance (approximately 0.1 Å) than the X-ray counterpart. While correcting the discrepancy would require

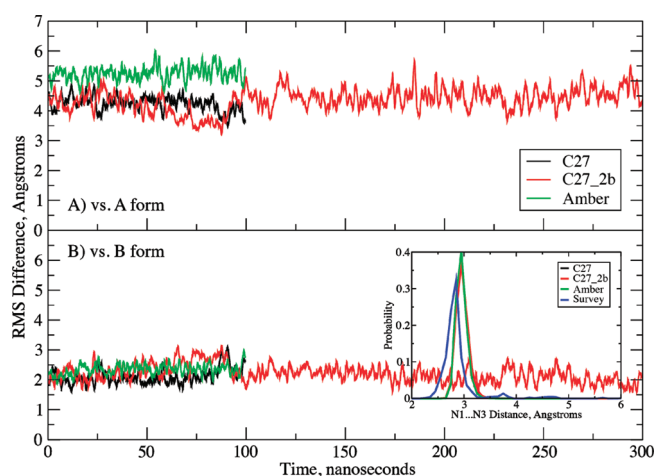


Figure 3. RMS difference versus time for the EcoRI dodecamer in solution. RMS differences vs the (A) canonical A form and (B) canonical B form of DNA for all non-hydrogen atoms in the nonterminal residues. Results are for the C27 (black), C27_2b (red), and AMBER Parm99bsc0 (green) force fields. Inset: Watson–Crick base pair interaction based on the $N1 \cdots N3$ distance distributions for the three force fields and data from the survey of B form DNA crystal structures (blue).

reoptimization of the base nonbonded parameters, a far from trivial task, it should be considered in future FF development efforts. It is worth noting that the $N1 \cdots N3$ distance in crystal simulations using a polarizable FF of the bases developed in our laboratory has a maximum approximately 0.1 Å shorter than that of C27,⁶¹ such that this issue appears resolved in the polarizable model.

Since the primary goal of the present effort was increased sampling of the BII state, the populations of the BI and BII states were determined over all nucleotides in the simulated systems (Table 7). In the analysis, the BI and BII states are defined as the difference between the ϵ and ζ dihedrals, where $BI = \epsilon - \zeta \leq 0^\circ$ (peak around -90°) and $BII = \epsilon - \zeta > 0^\circ$ (peak around $+90^\circ$). C27_2b shows an increase in sampling of the BII states over C27 in all cases, with the exception of 1ZF1 in ethanol where both FFs only sample a small amount of BII. The amount of BII in *EcoRI* and JunFos using AMBER is significantly less than with the C27_2b model; similar results are seen for a previously reported simulation²¹ of the sequence targeted by the Ndt80 transcription performed with the AMBER FF94.⁵

While C27_2b achieved an overall increased sampling of the BII conformation, the ability to properly treat the sampling of BII as a function of sequence is of special interest, and a more difficult objective. For three of the studied sequences, *EcoRI*, JunFos, and NF- κ b, experimental data are available on the percent BII as a function of sequence.^{20,24,62} For *EcoRI*, consistent with the data in Table 7, C27_2b yields increased BII sampling (Figure 4 and Table 8) in better agreement with experimental results as compared to both C27 and AMBER, although the amount of BII is systematically underestimated, a point addressed below. With respect to the base-step specific percent BII, C27_2b offers significant improvement over both C27 and AMBER (Figure 4 and Table 8). For the base-step specific percent BII, the correlation coefficients between simulation and experiment were 0.69 with C27_2b, 0.47 with C27, and 0.31 with AMBER. Results for JunFos and NF- κ b were similar to those for *EcoRI* (Table S2, Supporting Information). For JunFos, the average difference in

Table 7. Populations of the BI and BII States for the C27_2b Validation Simulations^a

DNA	C27		C27_2b		Amber	
	BI	BII	BI	BII	BI	BII
<i>EcoRI</i> ¹	0.89	0.11	0.75	0.25	0.82	0.18
GTAC2	0.97	0.03	0.92	0.08		
BDJ025	0.83	0.17	0.62	0.38		
1AXP	0.91	0.09	0.70	0.30		
1ZF1(H ₂ O)	0.85	0.15	0.54	0.46		
1ZF1(EtOH)	0.90	0.10	0.91	0.09		
1ZF7	0.81	0.19	0.55	0.45		
3BSE	0.87	0.13	0.70	0.30		
JunFos	0.92	0.08	0.76	0.24	0.90	0.10
NF- κ B	0.93	0.07	0.72	0.28		
Ndt80			0.76	0.24	0.83 ²	0.17 ²

^a Results obtained over 100 ns simulations. (1) Statistical analysis for *EcoRI* based on five 20 ns blocks from which the averages and standard deviations were obtained as follows (%BII, average \pm standard deviation): C27, 10.9 ± 1.2 ; C27_2b, 24.7 ± 3.2 ; and Amber, $18.2 \pm 1.9\%$. A t test shows that the difference between C27 and C27_2b BII populations is statistically significant, with a P value of <0.0001 . (2) Trajectory from Hart et al.²¹ where the Amber parm94 parameters⁵ were used.

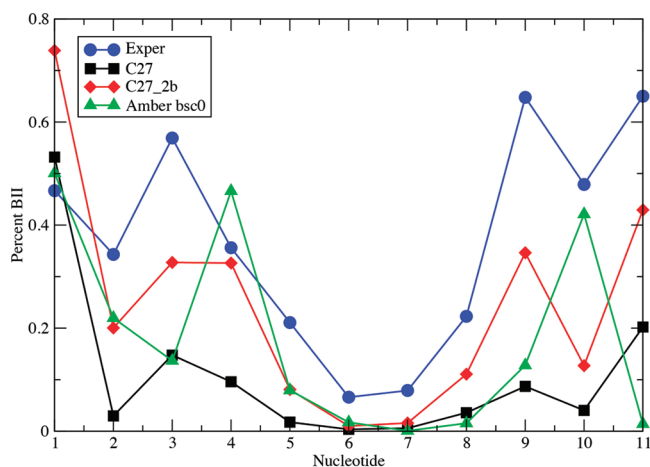


Figure 4. Percent BII conformation as a function of nucleotide for *EcoRI* from experiment and from the MD simulations using the C27, C27_2b and AMBER Parm99bsc0 force fields. Data for the symmetrically related basepair steps have been combined.

BII population between simulation and experiment was -7 for C27_2b compared to values of -22 and -20 for C27 and AMBER, respectively. The percent BII as a function of base step in JunFos was also improved with C27_2b with a correlation coefficient of 0.49 versus 0.30 and 0.26 for C27 and AMBER, respectively. In the case of NF- κ b (Table S2B), the average differences were -2 and -22 , and the correlation coefficients were 0.45 and 0.29 for C27_2b and C27, respectively. Thus, the reoptimization of the selected dihedral parameters increased the sampling of BII in a sequence-specific manner that is largely consistent with experimental observations.

The C27_2b FF clearly is an improvement over C27 and AMBER Parm99bsc0 with respect to sampling of the BII

Table 8. Average Percent BII as a Function of Base-Step for EcoRI, from Experiments and Simulations^a

base step	exptl	C27		C27_2b		Amber bsc0	
		avg	diff	avg	diff	avg	diff
C ₁ pG ₂	46.7	53.2 ± 4.7	6.5	73.9 ± 8.1	27.2	50.1 ± 13.9	3.4
G ₂ pC ₃	34.3	3.0 ± 1.3	-31.3	20.0 ± 10.6	-14.3	22.0 ± 5.8	-12.3
C ₃ pG ₄	56.9	14.8 ± 1.8	-42.1	32.8 ± 7.8	-24.2	13.7 ± 5.1	-43.2
G ₄ pA ₅	35.6	9.6 ± 0.8	-26.0	32.6 ± 6.5	-3.0	46.6 ± 7.1	11.0
A ₅ pA ₆	21.1	1.8 ± 0.4	-19.3	8.1 ± 1.6	-13.0	8.0 ± 2.1	-13.2
A ₆ pT ₇	6.6	0.4 ± 0.2	-6.2	1.0 ± 0.3	-5.6	1.7 ± 0.9	-4.9
T ₇ pT ₈	7.9	0.6 ± 0.2	-7.3	1.6 ± 0.5	-6.3	0.1 ± 0.1	-7.8
T ₈ pC ₉	22.3	3.6 ± 0.6	-18.7	11.1 ± 1.4	-11.2	1.6 ± 0.3	-20.8
C ₉ pG ₁₀	64.8	8.7 ± 0.5	-56.1	34.6 ± 6.1	-30.2	12.8 ± 3.8	-52.0
G ₁₀ pC ₁₁	47.9	4.1 ± 1.6	-43.9	12.7 ± 3.8	-35.2	42.1 ± 11.3	-5.8
C ₁₁ pG ₁₂	65.0	20.2 ± 14.5	-44.8	42.9 ± 5.9	-22.1	1.4 ± 0.5	-63.6
average difference			-26.3 ± 5.8		-12.5 ± 5.1		-19.0 ± 7.2
correlation			0.47		0.69		0.31

^a Results obtained over 100 ns simulations. Statistical analysis for the individual base steps accounting for the symmetry of the sequence based on five 20 ns blocks from which the averages and standard deviations were calculated. The errors for the average differences are the standard error over all of the base steps. Correlations are between the experimental and average simulation values over the base steps. Experimental data from ref 24 at 297.2 K.

conformation as compared to experimental results. However, for both *EcoRI* and *JunFos*, C27_2b still underestimates the extent of BII compared to experimental results. While this may be a limitation of the FF, the method of analysis may contribute to the difference. Analysis of the simulations was based on direct counting of the amount of BI and BII (i.e., based on BI is $\varepsilon - \zeta < 0$) from which the relative probabilities of the two states were obtained. Alternatively, in the ³¹P NMR analysis,²⁰ the chemical shift is converted to an average $\varepsilon - \zeta$ value that is used to identify the percent BI by interpolation between $\varepsilon - \zeta = 90^\circ$ (0% BI) and $\varepsilon - \zeta = -90^\circ$ (100% BI). While the results are similar for the two analyses, the more approximate interpolation method tends to overestimate the amount of BII (Figure S2 of the Supporting Information). Accordingly, the interpolation method used to estimate the BI/BII content from ³¹P NMR chemical shifts slightly overestimates the actual BII content. For example, with *JunFos*, the BII content when calculated using the interpolation method for C27_2b is in excellent agreement with the experimental estimate; the average difference between calculated and experimental percent BII is 2.3 although the correlation is slightly worse (0.43 vs 0.49).

As discussed above, in a 10 ns MD simulation of the Ndt80–DNA complex, the BII conformation at specific base steps was not maintained using C27.²¹ In the crystal structure of the complex, the BII conformation occurs at two important YpG base steps where arginines hydrogen-bond specifically with guanines; these were maintained with the AMBER Parm94 FF, leading to that FF being used in that study. In the crystal structure, the BII conformation is less pronounced for the T6'–G5' step ($\varepsilon - \zeta = 38^\circ$) than for the T4'–G3' step ($\varepsilon - \zeta = 84^\circ$). In simulations of the Ndt80–DNA complex, Parm94 gave 49% and 55% BII for these two steps, whereas the corresponding BII populations were 10% and 85%, respectively, in a 100 ns simulation with C27_2b done as part of the present study. The arginine hydrogen bonding patterns were similar with the two force fields, with R111 and R177 forming classical arginine double hydrogen bonds to the guanine N7 and O6 atoms. Both Parm94 and C27_2b formed the R111-Gua3' hydrogen bonds

>90% of the time, and the R177–Gua5' hydrogen bonds >95% of the time. Thus, C27_2b has rectified the problem with C27 observed by Hart et al.²¹

Alteration of the ε , ζ , and sugar dihedral parameters may also impact the flexibility of the FF, making it necessary to test this aspect of the model in MD simulations. This was addressed by analyzing the RMS fluctuations for selected systems and comparison with the NMR order parameters for *EcoRI*. Figure 5 presents the RMS fluctuations as a function of nucleotide for *EcoRI* for C27, C27_2b, and AMBER from 100 ns simulations. The overall pattern of fluctuations is similar for the three FFs, though the terminal base pairs are clearly more mobile in C27 and C27_2b as compared to AMBER. Despite more flexible termini, the fluctuations of the central nucleotides are similar for the three FFs, with slightly larger RMS fluctuations with C27_2b than with C27. The RMS fluctuations for C27 and C27_2b for 1ZF7 and 3BSE (Figure S3 of the Supporting Information) confirm the results from *EcoRI* (Figure 5), with C27_2b being more flexible than C27, a somewhat expected outcome considering the increased BII sampling with C27_2b.

Dynamics of the FFs were also tested on the basis of the reproduction of NMR order parameters for *EcoRI*⁶³ for the C27, C27_2b, and AMBER FFs. Results for the C1', C3', and C6/C8 atoms are in Table 9 for the individual palindromic strands in the duplex, to probe convergence of the results. In general, the C27_2b order parameters are lower than those for both C27 and AMBER, as seen in the average differences in Table 9, with C27 and AMBER having similar values and being in good overall agreement with experimental results. Thus, the increased fluctuations observed for C27_2b appear associated with order parameters systematically lower than those experimentally derived, though it may be related to the salt concentration in the simulations, as discussed below.

Correlation coefficients between the MD-based and NMR order parameters helped analyze how the FFs reflect those values as a function of sequence (Table 9). For C1', the C27 and C27_2b correlations range from 0.84 to 0.90 for all nucleotides (0.57–0.74 without terminal nucleotides). These values are

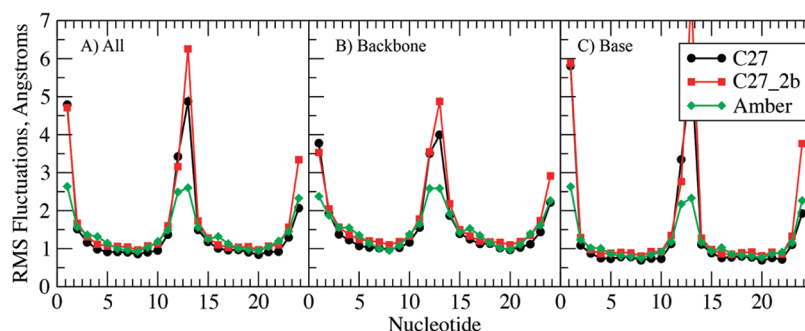


Figure 5. RMS Fluctuations of the EcoRI dodecamer as a function of nucleotide from 100 ns MD simulations using the C27 (black), C27_2b (red), and AMBER Parm99bsc0 (green) for the non-hydrogen atoms for the (A) full duplexes, (B) the phosphodiester backbone, or (C) the bases. Structures were least-squares aligned to the starting structures based on all non-hydrogen atoms prior calculation of the RMS fluctuations. The sequence numbering on the x axis covers the two strands of the oligonucleotide, with positions 12 and 13 corresponding to termini nucleotides.

higher than their AMBER counterpart. Similar levels of correlation are obtained for the C3'–H vector with all three FFs. For the C6/C8 values, the correlations are again similar for all the FFs. When the terminal nucleotides are excluded, the correlation coefficients are relatively small, possibly due to the small range of S^2 values. With both CHARMM FFs, the values are positive, while with AMBER they are negative, indicating slightly anti-correlated behavior. These results are generally consonant with the BII analysis as a function of sequence (Table 8 and Table S2), suggesting that C27_2b satisfactorily treats the analyzed properties as a function of sequence.

Due to their polyanionic nature, oligonucleotides are sensitive to salt concentration.^{60,64} To test the potential impact of salt on the reproduction of the NMR order parameters, a second C27_2b simulation of EcoRI was performed with the salt concentration adjusted to approximate the experimental regimen.⁶³ Interestingly, all of the calculated S^2 values tend to be larger in that simulation (Table S3), leading to average differences that are less negative than those obtained with C27_2b in low (i.e., neutralizing) salt (Table 9). The increase in the order parameters is largest at the termini, though increases in the duplex central region are evident. Thus, even a small change in salt concentration (i.e., five additional Na^+ and five additional Cl^- ions added to a simulation box of almost 26 000 atoms) appear to impact the calculated order parameters. These results, along with those for the method used to estimate the BI/BII ratio, indicate the difficulties and care that must be taken in performing a rigorous, quantitative comparison of computed and experimental data. Incidentally, experimentally determined BI/BII ratios have been shown to depend on the concentration of monovalent cations.²⁹

The above analyses indicate that C27_2b provides a reasonable treatment of the equilibrium between the A and B forms of DNA and improves the representation of the equilibrium between the BI and BII states. Yet, it was necessary to check that the C27_2b FF satisfactorily treated other detailed (but crucial) aspects of DNA structure. This involved examination of the probability distributions of dihedrals in the phosphodiester backbone, of the glycosidic linkage, of the sugar puckering, and of helicoidal parameters. Those distributions may be compared to X-ray survey data of DNA structures (see the Methods). In addition, time series may be analyzed to understand details of the dynamics of the systems. Comparison to crystallographic data is informative, especially for localized geometric properties. However, one must keep in mind that the sequence-specific properties

of DNA in solution may depart from crystallographic averages, especially regarding helicoidal parameters. Unfortunately, accurate structures of DNA duplexes in solution are still too scarce¹³ to provide sufficient reference data for systematic evaluation of simulated structure across sequences.

Dihedral and pucker distributions for the EcoRI dodecamer for the C27, C27_2b, and AMBER FFs are shown in Figure 6 along with NDB survey data obtained as part of the present study. Similar distributions for 3BSE, 1ZF7, and 1ZF1 in ethanol are included in Figure S4 of the Supporting Information. In EcoRI (Figure 6), the CHARMM force fields satisfactorily reproduce the NDB distributions with respect to both the location and ranges. With all FFs, there is a slight shift in the γ distribution toward higher values. In the glycosidic χ distribution, the conventional B form range from 140 to 270 is populated as expected, but a peak at higher χ around 280° is not present in the FF results. This peak in the crystal distribution is associated with terminal nucleotides in the survey; the terminal nucleotides were not included in the analysis of the MD simulations. Also, the CHARMM FFs show a minor but distinct population for χ around 210°, representing A-like χ conformations, which is slightly more stabilized in C27_2b than C27. This is consistent with more sampling of δ around 80° (associated with north sugar puckers) in C27_2b than in C27. Importantly, this increased sampling of the A-like χ conformation and sugar pucker has not destabilized the overall B form in C27_2b. With ϵ and ζ , C27_2b allows increased sampling in the regions of 240° and 165°, respectively, associated with the increased BII populations. The increased BII conformation is also associated with the shoulder in the β distribution in the region of 140°, consistent with the previously reported correlation between β and the BI and BII states.²⁷

Changes in sugar dihedral parameters (Table S1) altered the sugar pucker distributions while lowering the energy of the north conformation in model compound 2 (Table 3). Both CHARMM FFs reproduce the overall distribution in the south region (~140 to 180°, ~C2' endo pucker) associated with B form DNA, although the distinct peak at around 150° in the survey distribution is not reflected in the FF distributions. The shape of the distributions reproduces the shoulder in the vicinity of 120°, followed by low sampling at lower pucker values (~90°, associated with the east energy barrier), followed by a peak at 15° associated with north puckering common to A form DNA (~C3' endo). Interestingly, the NDB survey shows no sampling in the north region, although sampling is observed in the vicinity

Table 9. Order Parameters, S^2 , for EcoRI from NMR Experiments and MD Simulations^a

C1' atom	exptl		C27		C27_2b		Amber	
	S^2	SD	s1	s2	s1	s2	s1	s2
base								
1	0.52	0.02	0.26	0.30	0.17	0.02	0.56	0.38
2	0.78	0.03	0.80	0.80	0.78	0.59	0.70	0.78
3	0.74	0.03	0.66	0.65	0.54	0.61	0.66	0.72
4	0.88	0.03	0.85	0.86	0.79	0.82	0.67	0.70
5	0.84	0.03	0.89	0.88	0.82	0.83	0.74	0.76
6	N.A.	N.A.	0.77	0.79	0.57	0.57	0.81	0.82
7	0.92	0.02	0.74	0.76	0.66	0.66	0.85	0.85
8	0.86	0.02	0.83	0.82	0.81	0.80	0.82	0.82
9	0.68	0.03	0.64	0.61	0.56	0.54	0.71	0.69
10	0.85	0.02	0.80	0.84	0.59	0.81	0.73	0.67
11	0.71	0.02	0.66	0.78	0.62	0.64	0.78	0.77
12	N.A.	N.A.	0.30	0.69	0.23	0.57	0.50	0.67

difference analysis

average all			-0.06	-0.05	-0.14	-0.14	-0.05	-0.06
correlation			0.90	0.88	0.84	0.89	0.70	0.79
average_non_terminal			-0.04	-0.03	-0.12	-0.11	-0.06	-0.05
correlation			0.74	0.69	0.57	0.73	0.39	0.36

C3' atom	exptl		C27		C27_2b		Amber	
	S^2	SD	s1	s2	s1	s2	s1	s2
base								
1	0.39	0.02	0.29	0.35	0.15	0.07	0.52	0.36
2	N.A.	N.A.	0.75	0.79	0.78	0.59	0.79	0.78
3	N.A.	N.A.	0.57	0.53	0.45	0.54	0.70	0.65
4	N.A.	N.A.	0.85	0.85	0.79	0.84	0.74	0.78
5	0.90	0.02	0.90	0.87	0.80	0.81	0.78	0.78
6	0.79	0.03	0.67	0.71	0.35	0.37	0.82	0.83
7	N.A.	N.A.	0.63	0.71	0.44	0.44	0.83	0.82
8	0.79	0.03	0.81	0.78	0.76	0.73	0.79	0.78
9	0.67	0.04	0.46	0.42	0.45	0.39	0.66	0.67
10	N.A.	N.A.	0.73	0.80	0.37	0.78	0.80	0.75
11	N.A.	N.A.	0.39	0.72	0.35	0.52	0.74	0.79
12	0.43	0.05	0.49	0.70	0.35	0.48	0.54	0.61

difference analysis

average all			-0.06	-0.02	-0.18	-0.18	0.03	0.01
correlation			0.89	0.67	0.80	0.74	0.95	0.88
average_nonterminal			-0.08	-0.09	-0.20	-0.21	-0.02	-0.02
correlation			0.93	0.94	0.61	0.72	0.70	0.72

C6/C8 atoms	exptl		C27		C27_2b		Amber	
	S^2	SD	s1	s2	s1	s2	s1	s2
base								
1_C6	0.77	0.04	0.39	0.45	0.29	0.11	0.58	0.56
2_C8	0.81	0.07	0.87	0.87	0.84	0.84	0.86	0.85
3_C6	0.92	0.04	0.86	0.86	0.84	0.85	0.85	0.86
4_C8	N.A.	N.A.	0.90	0.90	0.88	0.88	0.88	0.88
5_C8	N.A.	N.A.	0.90	0.90	0.89	0.89	0.88	0.89
6_C8	N.A.	N.A.	0.91	0.91	0.90	0.90	0.91	0.91
7_C6	0.83	0.02	0.86	0.87	0.86	0.86	0.92	0.92
8_C6	0.87	0.04	0.86	0.86	0.85	0.85	0.91	0.91
9_C6	0.79	0.06	0.87	0.86	0.83	0.84	0.87	0.87

Table 9. Continued

C6/C8 atoms	exptl		C27		C27_2b		Amber	
	S^2	SD	s1	s2	s1	s2	s1	s2
base								
10_C8	0.88	0.04	0.89	0.90	0.87	0.88	0.88	0.87
11_C6	0.88	0.04	0.86	0.85	0.82	0.82	0.87	0.87
12_C8	0.91	0.08	0.70	0.80	0.61	0.65	0.76	0.78

difference analysis

average all			-0.05	-0.04	-0.09	-0.11	-0.02	-0.02
correlation			0.42	0.50	0.36	0.46	0.35	0.41
average_nonterminal			-0.01	0.00	-0.05	-0.04	0.00	0.01
correlation			0.09	0.06	0.12	0.18	-0.13	-0.09

^aResults from the simulations are presented individually for strand 1 (s1) and strand 2 (s2). Experimental data from Duchardt et al.⁶³ Analysis over the 5–100 ns portions of the trajectories. SD indicates the standard deviation in the experimental values. Difference and correlation coefficient calculated over nucleotides for which experimental data are available, excluding the terminal nucleotides.

of 90° for δ in the survey. Detailed analysis indicates that subtle differences in the nature of the sugar puckering (i.e., associated with differences in the five furanose ring dihedrals that define the pucker) in A vs B form crystal structures when δ is \sim 90° are present, leading to the lack of north sampling in the sugar phase in the survey data. Studies to better elucidate this effect are ongoing.

Concerning systems 1ZF7, 3BSE, and 1ZF1 in 75% ethanol (Figure S4), the results from 1ZF7 and 3BSE are similar to those for EcoRI, consistent with those structures sampling the B form in solution. With 1ZF7, increased sampling of the BII conformation, consistent with its high GC content (Table 1), is seen in the β , ϵ , and ζ distributions for C27_2b. In the 3BSE simulation, the additional peak in the ζ distribution with C27 and the additional sampling of north sugars in C27_2b are consistent with the reported disordered nature of the duplex in solution.⁶⁵ In particular, the authors note a six base A + T rich segment at the center of the strand, showing “unusually weak electron density, suggesting conformational fluctuations,” and propose that its disorder is “intrinsic to its sequence”. In this region, C27 gives an average RMS fluctuation of 1.64 Å over all residues, with C27_2b giving an average RMS fluctuation of 1.87 Å. For comparison, in the B form structures EcoRI and 1ZF7, and considering all nonterminal residues, C27 gives average RMS fluctuations of 1.24 and 1.20 Å, respectively, while C27_2b gives average RMS fluctuations of 1.43 and 1.49 Å, respectively.

For 1ZF1 in ethanol, where the survey data are that from A form helices, the difference in sampling associated with C27 assuming a B conformation while C27_2b remained in the A conformation is evident. The A form conformation in the C27_2b simulation nicely reproduces all of the survey distributions. Increased sampling of $\delta = \sim$ 80° and north pucker is evident. The model reproduces the shift in the location of the maxima in the ϵ and ζ distributions, versus that occurring in the B form survey data (i.e., compare survey results in Figure S4b and c), further indicating the conformational properties of C27_2b to be sensitive to changes in the environment.

The dihedral and pucker distributions with AMBER for EcoRI (Figure 6) are generally similar to the CHARMM FFs, but with notable differences. The AMBER α distribution is shifted to lower values compared to the survey, though the peak height is in good agreement. The peak height is also in good agreement for β ,

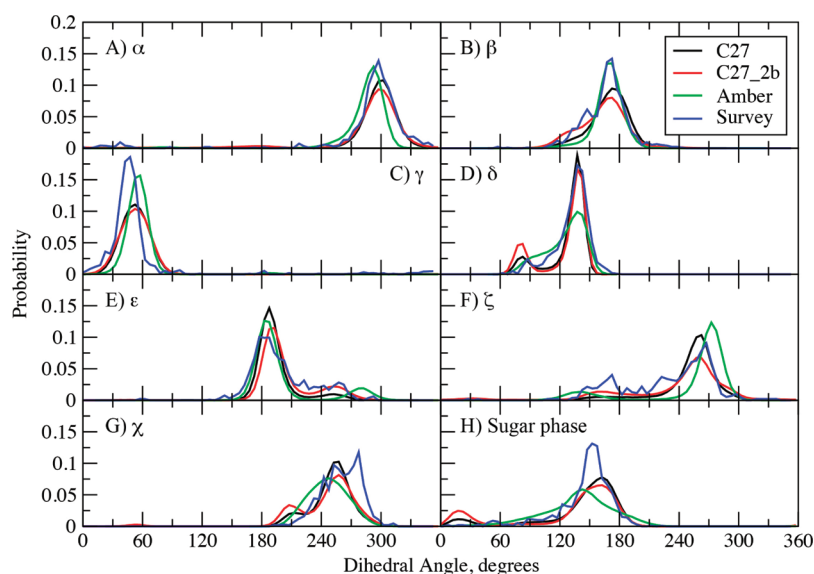


Figure 6. Dihedral angle and pseudorotation angle probability distributions from 100 ns MD simulations of *EcoRI* using the C27 (black), C27_2b (red), and AMBER Parm99bsc0 (green). Included are corresponding distributions from a NDB survey of all B form structures with a resolution ≤ 2.5 Å.

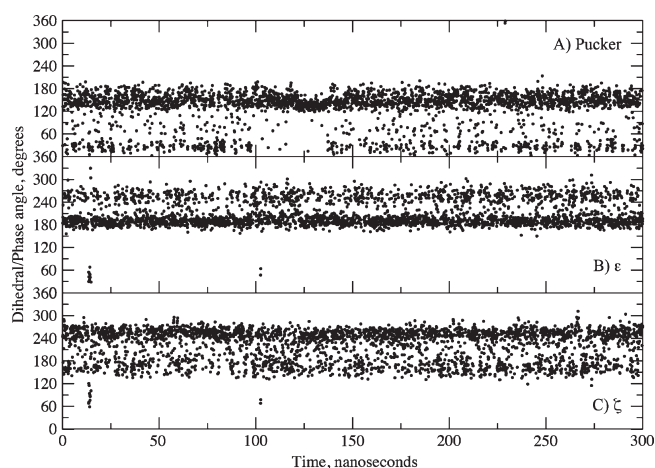


Figure 7. Time series from *EcoRI* C27_2b simulation for (A) sugar puckering of strand 1, nucleotide 3, (B) ϵ of strand 1, nucleotide 4, and (C) ζ of strand 1, nucleotide 4. Data points are shown for every 100 ps. See Figures S6, S7, and S8 of the Supporting Information for all puckering, ϵ and ζ time series. Note that in the BI state ϵ and ζ are approximately 190° and 270° , respectively, and in the BII state they are approximately 270° and 180° , respectively.

though the shoulder in the vicinity of 140° is too small, consistent with the lower amount of sampling of the BII conformation. A shift to higher values is present in the γ distribution as compared to that from the survey. With δ , the range covered by AMBER is similar to that from the survey, while sampling of the $\sim 90^\circ$ region is more a shoulder than a distinct peak. This is consistent with the sugar phase distribution, which is quite broad with no sampling of the north conformation. A study by Kollman and co-workers addressed this issue,⁶ though the resulting parametrization is not included in the AMBER parm99bsc0 FF. With ϵ and ζ , sampling of the 240° and 165° regions, respectively, is underestimated relative to that in the survey, again consistent with the lower amount of BII sampling. The major ϵ peak agrees well with experimental results, though the peak at higher values associated

with the BII conformation is systematically shifted relative to that from the survey. The major peak for ζ in the 270° region is shifted to values higher than in the survey, while the BII associated peak extends to values lower than in the survey. Finally, with χ , a broad range of values is sampled in general agreement with the survey data; a distinct peak at $\sim 210^\circ$, associated with A-like conformations, that occurs in the CHARMM force fields, is not present. This analysis indicates AMBER to generally sample ranges of the dihedrals seen in the survey data, consistent with the B form of the dodecamer being stable in solution; however, systematic shifts with respect to the survey data are present in a number of cases.

To better understand the nature of the conformational transitions giving rise to the distinct peaks in the distributions of the sugar pucker, ϵ and ζ (Figure 6), the corresponding time series from the C27_2b *EcoRI* simulation were analyzed. Selected time series are shown in Figure 7 with all of the pucker, ϵ , and ζ time series shown in Figures S5, S6, and S7 of the Supporting Information, respectively. These plots show that the three degrees of freedom have undergone a large number of transitions, such that the lifetimes of the distinct conformations are on the 10 ps time scale. Similar results were obtained in the AMBER *EcoRI* simulation (not shown). The relatively short lifetimes are consistent with NMR ^{13}C spin relaxation experiments from which relaxation times in sub-100-ps range were measured for the $\text{C1}'\text{--H1}'$ and $\text{C3}'\text{--H3}'$ vectors.⁶³ Concerning ϵ and ζ , some ^{31}P NMR experiments suggested that the free energy barrier for the BI to BII transition might be in the range of 12 to 15 kcal/mol.²⁴ This barrier height would suggest a lifetime on the order of milliseconds, significantly longer than observed with the present FFs. However, those estimates are based on the assumption of a two state model, which the present calculations indicate may not be appropriate given the significant sampling of intermediate states in both ϵ and ζ .

The “intrinsic” contribution to the barrier between the BI and BII states had been found to be less than 2.0 kcal/mol in QM calculations based on compound 2 without the base.³¹ Here, additional 2D ϵ vs ζ energy surfaces were obtained with C27, C27_2b, and QM at the MP2/6-31+G(d) level for model

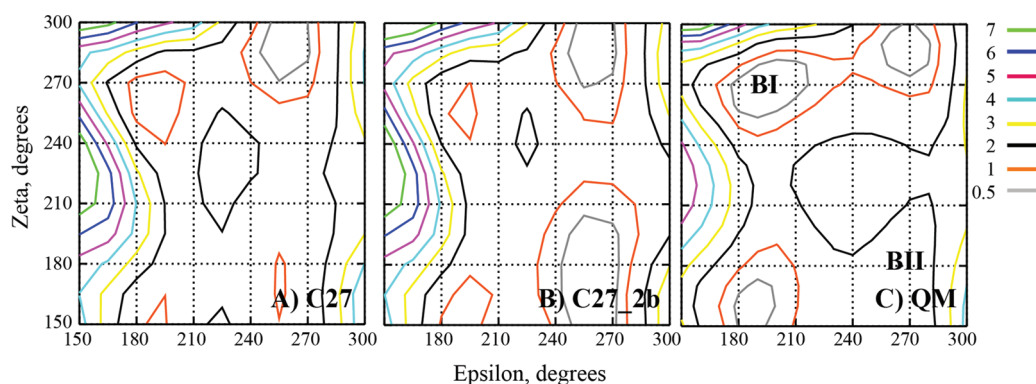


Figure 8. 2D potential energy surfaces of ϵ vs ζ for the (A) C27 FF, the (B) C27_2b FF, and (C) QM MP2/6-31+G(d) levels of theory for model compound **2** with the base omitted. Energies in kcal/mol. Sugar pucker was restrained to the south pucker by constraining $C1'-O4'-C4'-C3' = 0.0$, and the α dihedral was constrained to 300° .

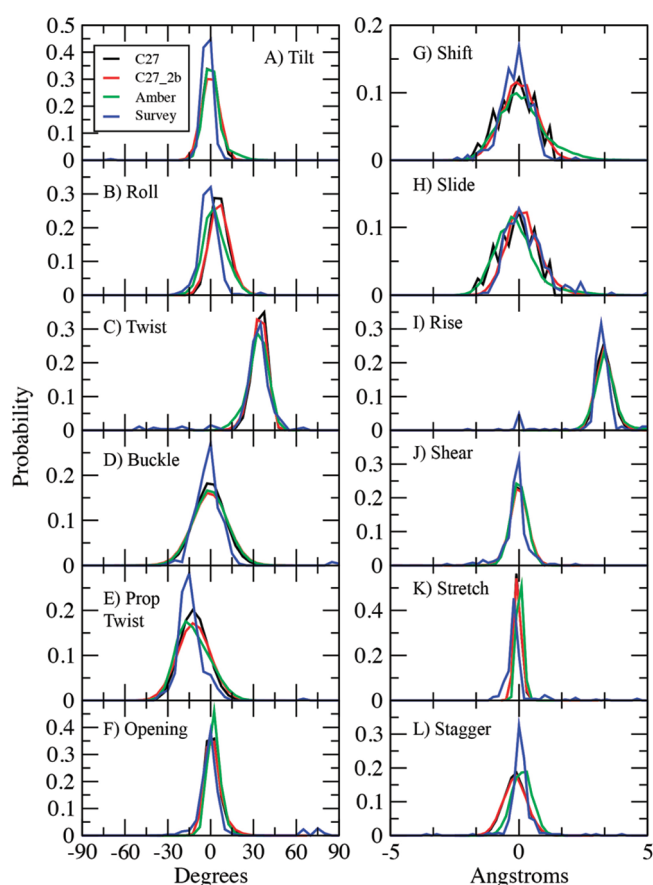


Figure 9. Helicoidal parameter probability distributions from 100 ns MD simulations using the C27 (black), C27_2b (red), and AMBER Parm99bsc0 (green). Included are corresponding distributions from a NDB survey of all B form structures with a resolution ≤ 2.5 Å.

compound **2** in which the base had been omitted. Shown in Figure 8 are the three energy surfaces with the locations of the BI and BII conformations shown on the right-hand panel. The overall shapes of the surfaces are similar, though differences are evident. The BI minimum is not as deep in the empirical surfaces versus the QM, while the opposite is true for the BII minimum. The increased depth of the BII minimum in C27_2b versus C27 is evident, consistent with the lowering of the relative BII energy

based on model compound **1** (Table 2). Concerning the barrier between the BI and BII conformations, there are two low energy paths along $\epsilon \sim 200$ and $\sim 270^\circ$ on all three surfaces. In all of the surfaces, the highest energies are between 1 and 2 kcal/mol with the barriers being lower in the empirical surfaces. A previous study estimated the free energy barrier to be in the range of 2.6 to 3.1 kcal/mol based on potential of mean force calculations between the A and B forms of DNA.⁶⁶ These values are consistent with the 1D surfaces for **1** (Figure 2) and indicate that the experimental estimate of >12 kcal/mol is not due to the intrinsic energies of the phosphodiester backbone, suggesting that more global structural phenomena (e.g., base stacking, sugar puckering etc.) may be contributing to the barrier if the model used to make the experimental estimates is appropriate. Future studies are required address this issue.

Helicoidal parameters are commonly used to define the orientations of the bases relative to the helical axis and vary among the different forms of DNA.^{51,52,67} Accordingly, the FFs should be able to reproduce experimental values for these descriptors. In the present study, the experimental data are probability distributions of selected helicoidal parameters based on the survey of crystallographic structures. Presented in Figure 9 are the survey data along with probability distributions from the *EcoRI* simulations using C27, C27_2b, and AMBER. In general, all three forces fields adequately reproduce the survey distributions. With roll, the AMBER distribution is in slightly better agreement with the survey while C27_2b yields slightly better overlap with the survey data for slide than both C27 and AMBER. Comparison of the C27 and C27_2b distributions show them to generally be similar. This is expected as the base parameters have not changed, although some correlations between backbone conformation and helicoidal parameters are known.

Analyses of selected helicoidal parameters as a function of nucleotide from the *EcoRI* simulations are presented in Figure S8 of the Supporting Information. The figures include the values from the *EcoRI* crystal structure 1BNA. All three FFs do well in reproducing the trends observed in the crystal structures. Notably, none of the FFs reproduce the large value of twist at base pair 2 (the first G along the X axis) and the small values of Roll and Rise at base pair 3. Concerning the overall values of twist, the C27, C27_2b, and AMBER values were 34.4 ± 4.8 , 33.8 ± 5.4 , and $32.8 \pm 6.4^\circ$, respectively, where the values are the averages and the standard deviations over the nucleotides. AMBER is in better agreement with the average value from the 1BNA crystal

structure of $32.8 \pm 10.6^\circ$, though both CHARMM force fields are in better agreement with the value for canonical B form DNA, 36° , consistent with previous studies.^{43,68}

Finally, to further test the generality of the C27_2b FF, it was used to simulate a Z-form dodecamer in its crystal environment (1LXJ, #9 in Table 1). RMSD analysis (Table 6) showed the simulations to stay close to the initial Z-form crystal structure. Dihedral distributions were also calculated and compared to survey data of Z-DNA structures that included unmodified and base-modified duplexes to get a satisfactory sample size (Figure S4d, Supporting Information). The simulated distributions are in overall reasonable agreement with the survey distributions. Notably, the two peaks associated with the syn and anti conformations about the glycosidic linkage (χ) are maintained, although some systematic shifts relative to experimental results are observed. Similar systematic shifts are seen in the calculated phase distributions as well as elsewhere, which is assumed to be associated with the FF being optimized to reproduce dihedral distributions of A and B forms of DNA. Despite these differences, it appears that C27_2b is of utility for simulations of Z DNA in the appropriate environment. While not shown, a simulation of 1LXJ in solution was unstable, with the DNA unwinding, indicating that C27_2b does not artificially stabilize Z DNA.

CONCLUSIONS

We have presented a refinement of the CHARMM27 all-atom additive force field for nucleic acids with emphasis on improving the sampling of the BII state of B form DNA. This required adjustment of selected parameters and validation of many facets of the resulting force field in relation to its representation of various forms of DNA and the conditions in which they are stable. Parameter optimization was initially based on model compounds allowing for systematic changes in only the dihedral parameters to improve agreement with target QM data. This yielded several parameter sets that were tested in condensed phase simulations of a training set of oligonucleotides. This procedure was also applied to improve treatment of the sugar pucker with respect to north and south conformations that dominate the A and B forms of DNA, respectively. The most promising FF model (C27_2b) was then used in simulations of other DNA molecules for a duration of 100 ns, with the *EcoRI* simulation extended to 300 ns. Analysis of these simulations showed the C27_2b FF model to reproduce a range of experimental data in DNA, thereby providing convincing validation for C27_2b. In particular, C27_2b provides a more accurate treatment of the BI/BII equilibrium in DNA, which is a significant improvement over C27. Simulations using the AMBER parm99bsc0 FF were also undertaken on *EcoRI* and JunFos, yielding overall results similar to those for both CHARMM force fields, although some differences in the sampling of backbone dihedrals were observed. While general agreement of the C27_2b with a range of experimental observables was obtained, technical difficulties in obtaining rigorous quantitative comparisons of calculated and experimental results for the BII populations and NMR order parameters were noted. Also, the dearth of accurate DNA structures in solution, with experimentally characterized populations and dynamics, remains a fundamental limitation for the development and optimization of DNA force fields. The selected DNA FF will be included with a recent revision of the CHARMM RNA FF,³⁴ yielding the CHARMM36 all-atom additive force field for nucleic acids.

ASSOCIATED CONTENT

S Supporting Information. Included are tables of the modified dihedral parameters, percent BII for JunFos and NF- κ b, and order parameter for *EcoRI* in high salt and figures of the RMS differences vs time for 1ZF1; BII content for JunFos; RMS fluctuations for 1ZF7 and 3BSE; dihedral and pucker probability distributions for 1ZF7, 3BSE, 1ZF1, and 1LXJ; and the pucker, ϵ , and ζ time series for all nucleotides in *EcoRI*. This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: Lennart.Nilsson@ki.se, alex@outerbanks.umaryland.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

NIH (GM051501) and the Swedish Research Council are thanked for financial support. We acknowledge the NSF Tera-Grid Computing, the Pittsburgh Supercomputing Center, and the Department of Defense High Performance Computing for their generous allocations of computer time. We thank Dr. Brigitte Hartmann for helpful discussions about DNA structure and the importance of its BII state.

REFERENCES

- (1) MacKerell, A. D., Jr. Empirical Force Fields for Biological Macromolecules: Overview and Issues. *J. Comput. Chem.* **2004**, *25*, 1584–1604.
- (2) Orozco, M.; Noy, A.; Pérez, A. Recent advances in the study of nucleic acid flexibility by molecular dynamics. *Curr. Opin. Struct. Biol.* **2008**, *18*, 185–193.
- (3) Perez, A.; Luque, F. J.; Orozco, M. Frontiers in Molecular Dynamics Simulations of DNA. *Acc. Chem. Res.* **2011**.
- (4) Foloppe, N.; MacKerell, A. D., Jr. All-atom empirical force field for nucleic acids: 1) Parameter optimization based on small molecule and condensed phase macromolecular target data. *J. Comput. Chem.* **2000**, *21*, 86–104.
- (5) 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*, 5179–5197.
- (6) Cheatham, T. E., III; Cieplak, P.; Kollman, P. A. A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct. Dyn.* **1999**, *16*, 845–861.
- (7) Langley, D. R. Molecular dynamics simulations of environment and sequence dependent DNA conformation: The development of the BMS nucleic acid force field and comparison with experimental results. *J. Biomol. Struct. Dyn.* **1998**, *16*, 487–509.
- (8) Soares, T. A.; Hunenberger, P. H.; Kastenholz, M. A.; Kraeutler, V.; Lenz, T.; Lins, R.; Oostenbrink, C.; van Gunsteren, W. An improved nucleic acid parameter set for the GROMOS force field. *J. Comput. Chem.* **2005**, *26*, 725–737.
- (9) MacKerell, A. D., Jr.; Nilsson, L. Theoretical studies of nucleic acids and nucleic acid-protein complexes using CHARMM. In *Computational Studies of DNA and RNA: Molecular dynamics, quantum chemistry, mesoscopic modeling*; Sponer, J., Lankas, F., Eds.; Springer: The Netherlands, 2006; pp 73–94.
- (10) Lavery, R.; Zakrzewska, K.; Beveridge, D.; Bishop, T. C.; Case, D. A.; Cheatham, T., 3rd; Dixit, S.; Jayaram, B.; Lankas, F.; Laughton, C.

A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. *Nucleic Acids Res.* **2010**, *38*, 299–313.

(11) Foloppe, N.; Nilsson, L. Toward a Full Characterization of Nucleic Acid Components in Aqueous Solution: Simulations of Nucleosides. *J. Phys. Chem. B* **2005**, *109*, 9119–9131.

(12) Zuo, X.; Cui, G.; Merz, K. M., Jr.; Zhang, L.; Lewis, F. D.; Tiede, D. M. X-ray diffraction "fingerprinting" of DNA structure in solution for quantitative evaluation of molecular dynamics simulation. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 3534–3539.

(13) Heddi, B.; Foloppe, N.; Oguey, C.; Hartmann, B. Importance of accurate DNA structures in solution: the Jun-Fos model. *J. Mol. Biol.* **2008**, *382*, 956–970.

(14) Isaacs, R. J.; Spielmann, H. P. Insight into G–T mismatch recognition using molecular dynamics with time-averaged restraints derived from NMR spectroscopy. *J. Am. Chem. Soc.* **2004**, *126*, 583–590.

(15) Pérez, A.; Marchán, I.; Svozil, D.; Sponer, J.; Cheatham, I.; Laughton, C. A.; Orozco, M. Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of alpha/gamma Conformers. *Biophys. J.* **2007**, *92*, 3817–3829.

(16) Joung, I. S.; Cheatham, T. E., 3rd. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B* **2008**, *112*, 9020–9041.

(17) Yildirim, I.; Stern, H. A.; Kennedy, S. D.; Tubbs, J. D.; Turner, D. H. Reparameterization of RNA χ Torsion Parameters for the AMBER Force Field and Comparison to NMR Spectra for Cytidine and Uridine. *J. Chem. Theory Comput.* **2010**, *6*, 1520–1531.

(18) Banas, P.; Hollas, D.; Zgarbova, M.; Jurecka, P.; Orozco, M.; Cheatham, I., T.E.; Sponer, J.; Otyepka, M. Performance of Molecular Mechanics Force Fields for RNA Simulations: Stability of UUCG and GNRA Hairpins. *J. Chem. Theory Comput.* **2010**, *6*, 3836–3849.

(19) Zgarbova, M.; Otyepka, M.; Sponer, J.; Mladek, A.; Banas, P.; Cheatham, T. E., 3rd.; Jurecka, P. Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. *J. Chem. Theory Comput.* **2011**, *7*, 2886–2902.

(20) Heddi, B.; Foloppe, N.; Bouchemal, N.; Hantz, E.; Hartmann, B. Quantification of DNA BI/BII backbone states in solution. Implications for DNA overall structure and recognition. *J. Am. Chem. Soc.* **2006**, *128*, 9170–9177.

(21) Hart, K.; Nilsson, L. Investigation of transcription factor Ndt80 affinity differences for wild type and mutant DNA: A molecular dynamics study. *Proteins: Struct., Funct., Bioinf.* **2008**, *73*, 325–337.

(22) Fratini, A. V.; Kopka, M. L.; Drew, H. R.; Dickerson, R. E. Reversible bending and helix geometry in a B-DNA dodecamer: CGCGAATTBrCGCG. *J. Biol. Chem.* **1982**, *257*, 14686–14707.

(23) Gorenstein, D. G.; David, M. J. L.; Dahlberg, J. E. 31P NMR of DNA. In *Methods Enzymol.*; Academic Press: Waltham, MA, 1992; pp 254–286.

(24) Tian, Y.; Kayatta, M.; Shultz, K.; Gonzalez, A.; Mueller, L. J.; Hatcher, M. E. 31P NMR Investigation of Backbone Dynamics in DNA Binding Sites. *J. Phys. Chem. B* **2008**, *113*, 2596–2603.

(25) Heddi, B.; Oguey, C.; Lavelle, C.; Foloppe, N.; Hartmann, B. Intrinsic flexibility of B-DNA: the experimental TRX scale. *Nucleic Acids Res.* **2010**, *38*, 1034–1047.

(26) Oguey, C.; Foloppe, N.; Hartmann, B. Understanding the Sequence-Dependence of DNA Groove Dimensions: Implications for DNA Interactions. *PLoS ONE* **2010**, *5*, e15931.

(27) Djuranovic, D.; Hartmann, B. Conformational Characteristics and Correlations in Crystal Structures of Nucleic Acid Oligonucleotides: Evidence for Sub-states. *J. Biomol. Struct. Dyn.* **2003**, *20*, 771–788.

(28) Lamoureux, J. S.; Stuart, D.; Tsang, R.; Wu, C.; Glover, J. N. M. Structure of the sporulation-specific transcription factor Ndt80 bound to DNA. *EMBO J.* **2002**, *21*, 5721–5732.

(29) Heddi, B.; Foloppe, N.; Hantz, E.; Hartmann, B. The DNA structure responds differently to physiological concentrations of K(+) or Na(+). *J. Mol. Biol.* **2007**, *368*, 1403–1411.

(30) Hartmann, B.; Piazzola, D.; Lavery, R. BI-BII transitions in B-DNA. *Nucl. Acid Res.* **1993**, *21*, 561–568.

(31) Foloppe, N.; MacKerell, A. D., Jr. Contribution of the Phosphodiester Backbone and Glycosyl Linkage Intrinsic Torsional Energetics to DNA Structure and Dynamics. *J. Phys. Chem. B* **1999**, *103*, 10955–10964.

(32) Lamoureux, J. S.; Maynes, J. T.; Glover, M. J. N. Recognition of 5'-YpG-3' Sequences by Coupled Stacking/Hydrogen Bonding Interactions with Amino Acid Residues. *J. Mol. Biol.* **2004**, *335*, 399–408.

(33) Svozil, D.; Kalina, J.; Omelka, M.; Schneider, B. DNA conformations and their sequence preferences. *Nucl. Acid Res.* **2008**, *36*, 3690–3706.

(34) Denning, E. J.; Priyakumar, U. D.; Nilsson, L.; MacKerell, A. D., Jr. Impact of 2'-hydroxyl sampling on the conformational properties of RNA: Update of the CHARMM all-atom additive force field for RNA. *J. Comput. Chem.* **2011**, *32*, 1929–1943.

(35) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A. *Gaussian 09*; Gaussian, Inc.: Wallingford, CT, 2009.

(36) Shao, Y.; Fusti-Molnar, L.; Jung, Y.; Kussmann, J.; Ochsenfeld, C.; Brown, S. T.; Gilbert, A. T. B.; Slipchenko, L. V.; Levchenko, S. V.; O'Neill, D. P. *Q-Chem*, 3.1 ed; Q-Chem, Inc.: Pittsburgh, PA, 2007.

(37) MacKerell, A. D., Jr. Contribution of the intrinsic mechanical energy of the phosphodiester linkage to the relative stability of the A, BI and BII forms of duplex DNA. *J. Phys. Chem. B* **2009**, *113*, 3235–3244.

(38) Berman, H. M.; Battistuz, T.; Bhat, T. N.; Bluhm, W. F.; Bourne, P. E.; Burkhardt, K.; Feng, Z.; Gilliland, G. L.; Iype, L.; Jain, S. The protein data bank. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2002**, *58*, 899–907.

(39) Berman, H. M.; Olson, W. K.; Beveridge, D. L.; Westbrook, J.; Gelbin, A.; Demeny, T.; Hsieh, S.-H.; Srinivasan, A. R.; Schneider, B. The Nucleic Acid Database: A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.* **1992**, *63*, 751–759.

(40) Foloppe, N.; Hartmann, B.; Nilsson, L.; MacKerell, A. D., Jr. Intrinsic Conformational Energetics Associated with the Glycosyl Torsion in DNA: a Quantum Mechanical Study. *Biophys. J.* **2002**, *82*, 1554–1569.

(41) Brooks, B. R.; Brooks, C. L., III; MacKerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S. CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **2009**, *30*, 1545–1614.

(42) 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.

(43) MacKerell, A. D., Jr.; Banavali, N. K. All-atom empirical force field for nucleic acids: 2) Application to solution MD simulations of DNA. *J. Comput. Chem.* **2000**, *21*, 105–120.

(44) 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.

(45) Hoover, W. G. Canonical Dynamics - Equilibrium Phase-Space Distributions. *Phys. Rev. A* **1985**, *31*, 1695–1697.

(46) Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, R. W. Constant Pressure Molecular Dynamics Simulation: The Langevin Piston Method. *J. Chem. Phys.* **1995**, *103*, 4613–4621.

(47) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.

(48) Nilsson, L. Efficient table lookup without inverse square roots for calculation of pair wise atomic interactions in classical simulations. *J. Comput. Chem.* **2009**, *30*, 1490–1498.

(49) Darden, T. A.; York, D.; Pedersen, L. G. Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(50) Steinbach, P. J.; Brooks, B. R. New Spherical-Cutoff Methods of Long-Range Forces in Macromolecular Simulations. *J. Comput. Chem.* **1994**, *15*, 667–683.

(51) Lavery, R.; Sklenar, H. The definition of generalized helicoidal parameters and of the axis of curvature for irregular nucleic acids. *J. Biomol. Str. Dyn.* **1988**, *6*, 63–91.

- (52) Ravishanker, G.; Swaminathan, S.; Beveridge, D. L.; Lavery, R.; Sklenar, H. Conformational and Helicoidal Analysis of 30 ps of Molecular Dynamics on the d(CGCGAATTCGCG) Double Helix: "Curves", Dials and Windows. *J. Biomol. Str. Dyn.* **1989**, *6*, 669–699.
- (53) Banavali, N. K.; MacKerell, A. D., Jr. Free Energy and Structural Pathways of Base Flipping in a DNA GCGC containing sequence. *J. Mol. Biol.* **2002**, *319*, 141–160.
- (54) Priyakumar, U. D.; MacKerell, A. D., Jr. Base Flipping in a GCGC Containing DNA Dodecamer: A Comparative Study of the Performance of the Nucleic Acid Force Fields, CHARMM, AMBER and BMS. *J. Chem. Theory Comput.* **2005**, *2*, 187–200.
- (55) Priyakumar, D.; MacKerell, A. D., Jr. Proton Exchange Experiments on Duplex DNA Primarily Monitor the Opening of Purine Bases. *J. Am. Chem. Soc.* **2006**, *678*–679.
- (56) Foloppe, N.; MacKerell, A. D., Jr. Intrinsic Conformational Properties of Deoxyribonucleosides: Implicated role for cytosine in the equilibrium between the A, B and Z forms of DNA. *Biophys. J.* **1999**, *76*, 3206–3218.
- (57) Srinivasan, J.; Withka, J. M.; Beveridge, D. L. Molecular dynamics of an in vacuo model of duplex d(CGCGAATTCGCG) in the B-form based on the amber 3.0 force field. *Biophys. J.* **1990**, *58*, 533–547.
- (58) McConnell, K. J.; Nirmala, R.; Young, M. A.; Ravishanker, G.; Beveridge, D. L. A Nanosecond Molecular Dynamics Trajectory for a B DNA Double Helix: Evidence for Substates. *J. Am. Chem. Soc.* **1994**, *116*, 4461–4462.
- (59) Sen, S.; Nilsson, L. Structure, Interactions, Dynamics and Solvent Effects on the DNA-EcoRI Complex in Aqueous Solution from MD Simulation. *Biophys. J.* **1999**, *77*, 1782–1800.
- (60) Manning, G. S. The molecular theory of polyelectrolyte solutions with applications to the electrostatic properties of polynucleotides. *Quart. Rev. Biophys.* **1978**, *11*, 179–246.
- (61) Baker, C. M.; Anisimov, V. M.; MacKerell, A. D., Jr. Development of CHARMM polarizable force field for nucleic acid bases based on the classical Drude oscillator model. *J. Phys. Chem. B* **2011**, *115*, 580–596.
- (62) Tisne, C.; Delepierre, M.; Hartmann, B. How NF-kappaB can be attracted by its cognate DNA. *J. Mol. Biol.* **1999**, *293*, 139–150.
- (63) Duchardt, E.; Nilsson, L.; Schleucher, J. Cytosine ribose flexibility in DNA: a combined NMR ¹³C spin relaxation and molecular dynamics simulation study. *Nucleic Acids Res.* **2008**, *36*, 4211–4219.
- (64) Record, M. T., Jr.; Anderson, C. F.; Lohman, T. M. Thermodynamic analysis of ion effects on the binding and conformational equilibria of proteins and nucleic acids: the roles of ion association or release, screening, and ion effects on water activity. *Quart. Rev. Biophys.* **1978**, *11*, 103–178.
- (65) Narayana, N.; Weiss, M. A. Crystallographic analysis of sex-specific enhancer element: sequence-dependent DNA structure, hydration and dynamics. *J. Mol. Biol.* **2009**, *385*, 469–490.
- (66) Banavali, N. K.; Roux, B. Free energy landscape of A-DNA to B-DNA conversion in aqueous solution. *J. Am. Chem. Soc.* **2005**, *127*, 6866–6876.
- (67) Dickerson, R. E. DNA bending: the prevalence of kinkiness and the virtues of normality. *Nucleic Acids Res.* **1998**, *26*, 1906–1926.
- (68) Reddy, S. Y.; Leclerc, F.; Karplus, M. DNA polymorphism: a comparison of force fields for nucleic acids. *Biophys. J.* **2003**, *84*, 1421–1449.
- (69) Langlois D'Estaintot, B.; Dautant, A.; Courseille, C.; Precigoux, G. Orthorhombic Crystal Structure of the A-DNA Octamer d(GTACGTAC). Comparison with the Tetragonal Structure. *Eur. J. Biochem.* **1993**, *213*, 673–682.
- (70) Grzeskowiak, K.; Yanagi, K.; Prive, G. G.; Dickerson, R. E. The Structure of B-Helical CGATCGATCG and Comparison with CCAACGTTGG. The Effect of Base Pair Reversals. *J. Biol. Chem.* **1991**, *266*, 8861–8883.
- (71) Drew, H. R.; Wing, R. M.; Takano, T.; Broka, C.; Tanaka, S.; Itakura, K.; Dickerson, R. S. Structure of a B-DNA dodecamer: Conformation and Dynamics. *Proc. Natl. Acad. Sci. U.S.A.* **1981**, *78*, 2179–2183.
- (72) Drew, H. R.; Dickerson, R. E. Structure of a B-DNA Dodecamer III. Geometry of Hydration. *J. Mol. Biol.* **1981**, *151*, 535–556.
- (73) Drew, H. R.; Samson, S.; Dickerson, R. E. Structure of a B-DNA Dodecamer at 16 K. *Proc. Natl. Acad. Sci. U.S.A.* **1982**, *79*, 4040–4044.
- (74) Holbrook, S. R.; Dickerson, R. E.; Kim, S.-H. Anisotropic Thermal-Parameter Refinement of the DNA Dodecamer CGCGAATTCGCG by the Segmented Rigid-Body Method. *Acta Crystallogr., Sect. B* **1985**, *41*, 255–262.
- (75) Westhof, E. Re-Refinement of the B-Dodecamer d(CGCGAATTCGCG) with a Comparative Analysis of the Solvent in it and in the Z-Hexamer d(5BrCG5BrCG5BrCG). *J. Biomol. Struct. Dyn.* **1987**, *8*, 581–600.
- (76) Johansson, E.; Parkinson, G.; Neidle, S. A new crystal form for the dodecamer C-G-C-G-A-A-T-T-C-G-C-G: symmetry effects on sequence-dependent DNA structure. *J. Mol. Biol.* **2000**, *300*, 551–561.
- (77) Zuo, X.; Tiede, D. M. Resolving conflicting crystallographic and NMR models for solution-state DNA with solution X-ray diffraction. *J. Am. Chem. Soc.* **2005**, *127*, 16–17.
- (78) Gyi, J. I.; Lane, A. N.; Conn, G. L.; Brown, T. Solution structures of DNA:RNA hybrids with purine-rich and pyrimidine-rich strands: comparison with the homologous DNA and RNA duplexes. *Biochemistry* **1998**, *37*, 73–80.
- (79) Hays, F. A.; Teegarden, A.; Jones, Z. J.; Harms, M.; Raup, D.; Watson, J.; Cavaliere, E.; Ho, P. S. How sequence defines structure: a crystallographic map of DNA structure and conformation. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 7157–7162.
- (80) Thiagarajan, S.; Sathesh Kumar, P.; Rajan, S. S.; Gautham, N. Structure of d(TGCGCA)₂ at 298K: comparison of the effects of sequence and temperature. *Acta Crystallogr., Sect. D* **2002**, *58*, 1381–1384.
- (81) Altona, C.; Sundaralingam, M. Conformational Analysis of the Sugar Ring in Nucleosides and Nucleotides. Improved Method for the Interpretation of Proton Magnetic Resonance Coupling Constants. *J. Am. Chem. Soc.* **1973**, *95*, 2333–2344.

50 Years of Lifson–Roig Models: Application to Molecular Simulation Data

Andreas Vitalis* and Amedeo Caflisch

Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

S Supporting Information

ABSTRACT: Simple helix–coil transition theories have been indispensable tools in the analysis of data reporting on the reversible folding of α -helical polypeptides. They provide a transferable means to not only characterize different systems but to also compare different techniques, viz., experimental probes monitoring helix–coil transitions in vitro or biomolecular force fields in silico. This article addresses several issues with the application of Lifson–Roig theory to helix–coil transition data. We use computer simulation to generate two sets of ensembles for the temperature-controlled, reversible folding of the 21-residue, alanine-rich FS peptide. Ensembles differ in the rigidity of backbone bond angles and are analyzed using two distinct descriptors of helicity. The analysis unmasks an underlying phase diagram that is surprisingly complex. The complexities give rise to fitted nucleation and propagation parameters that are difficult to interpret and that are inconsistent with the distribution of isolated residues in the α -helical basin. We show that enthalpies of helix formation are more robustly determined using van't Hoff analysis of simple measures of helicity rather than fitted propagation parameters. To overcome some of these issues, we design a simple variant of the Lifson–Roig model that recovers physical interpretability of the obtained parameters by allowing bundle formation to be described in simple fashion. The relevance of our results is discussed in relation to the applicability of Lifson–Roig models to both in silico and in vitro data.

INTRODUCTION

Elucidating the helix–coil transition microscopically has long been deemed to be of utmost importance for the understanding of protein folding, and the reader is referred to excellent review articles for further reading.^{1,2} The process is of such elementary nature that it has also become an indispensable benchmark for the development of biomolecular force fields.^{3–7}

Helix–coil transition data are often analyzed in an established statistical framework such as that of Zimm and Bragg,⁸ Gibbs and DiMarzio,⁹ or Lifson and Roig (LR).¹⁰ In the latter, it is assumed that the potential energy function of the system can be mapped to terms written over the ϕ/ψ angles of individual polypeptide residues with the exception of an α -helical hydrogen-bonding term coupling residue i energetically to residues $i - 1$ and $i + 1$. This term is triggered as soon as three consecutive residues are in a helix-competent conformation, and the resultant favorable energy contribution is mapped exclusively onto residue i . In the absence of hydrogen bonds, the statistical weights of helix-competent vs helix-incompetent (“coil”) states correspond to the respective, partial integrals over the Ramachandran map that due to the lack of residue–residue coupling can be formulated for each residue individually:

$$u'_i = \int_{c_i} e^{-\beta U(\phi, \psi)} d\phi_i d\psi_i \quad (1)$$

$$v'_i = \int_{h_i} e^{-\beta U(\phi, \psi)} d\phi_i d\psi_i \quad (2)$$

Here, c_i and h_i denote the helix-incompetent and helix-competent regions of ϕ/ψ space, respectively, while U is the (unknown) potential energy function. The LR model stipulates that whenever three consecutive residues are in helical conformation, stabilization occurs and another statistical weight, w'_i , is invoked. Recognizing the arbitrary absolute scale of the energy in

the system, the statistical weights can be normalized by u' . Then, the low level of coupling¹⁰ allows the partition function to be expressed in matrix form:

$$Z = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix} \cdot \left[\prod_{i=1}^{N_r} \begin{pmatrix} w_i & v_i & 0 \\ 0 & 0 & 1 \\ v_i & v_i & 1 \end{pmatrix} \right] \cdot \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \quad (3)$$

Here, N_r is the number of residues with peptide bonds on both sides and is equivalent to the number of amino acids for capped polypeptides. If we ignore any sequence specificity (including end effects), the matrices become identical, i.e., all residue subscripts can be dropped, and it is possible to obtain global averages as follows:

$$\begin{aligned} \langle N_h \rangle &= \frac{\partial \ln Z}{\partial \ln w} \\ \langle N_s \rangle &= \frac{\partial \ln Z}{\partial \ln v_{12}} \end{aligned} \quad (4)$$

Here, N_h denotes the number of α -helical hydrogen bonds, and N_s the number of helical segments. Matrix element v_{12} refers to a single instance of v in the matrix. Note that N_s by definition includes segments of only two helical residues in a row with no hydrogen bonds formed. This is because the formalism in eq 3 only scans three consecutive residues, and v_{12} corresponds to states of the type “hhc” (helix, helix, coil) regardless of configuration of preceding residues. Further illustrations of LR theory using a simple example are provided in the Supporting Information.

Received: October 24, 2011

Published: December 14, 2011

Because w denotes the ratio of the statistical weights of hydrogen-bonded and coil states for an individual residue, it is often assumed to correspond directly to the stepwise equilibrium constant of helix elongation, i.e., $-\beta^{-1} \ln w \approx \Delta G_{\text{hb}}$, where β is the inverse thermal energy, and ΔG_{hb} the free energy gain associated with the formation of a single hydrogen bond. It is possible to determine w from equilibrium experiments that are able to estimate helix content directly, such as temperature-dependent circular dichroism (CD) spectroscopy, by fitting the raw data to a two-state model that allows extraction of $\langle N_{\text{h}} \rangle$ and subsequent fitting to yield w . This requires knowledge of ν , which is often obtained independently or can be fit if data for multiple chain lengths are available.¹¹ Limitations of the LR model were established and analyzed soon after publication of the original model, and extensions were suggested.^{12,13} Throughout the past two decades, specific modifications were proposed that incorporate helix capping,¹⁴ short-range side chain interactions,¹⁵ or extensions beyond the triplet model.¹⁶

Experimental analyses of the helix–coil transition designed to extract more than just helix content and to interpret the results in terms of a microscopic theory have had to utilize assumptions to avoid overfitting the data. Rohl et al.¹⁷ used the kinetics of amide proton exchange to show that a single model with essentially three parameters can fit data at a single temperature for polypeptides of the series Ace-(AAKAA)_{*m*}-Y-NH₂ reasonably well for values of m ranging from 1 to 10. Such a simple model was obtained by assuming a homopolymer and by assuming that exchange in the only considered hydrogen-bonded state (α -helix) is completely quenched. Then, the free parameters were the exchange rate in the coil-state and the aforementioned helix nucleation and propagation parameters.

Later, the same authors showed that, for a nearly identical series of peptides and over a limited range of temperatures, two types of fits with similar quality could be obtained, both using a T-independent helix nucleation parameter and again assuming homopolymeric behavior.¹⁸ In the first fit, T-dependent propagation parameters were derived from CD, and the exchange rates in the coil-state were fitted, whereas in the second, the exchange rate was fixed to that observed for the shortest peptide, and helix propagation parameters were fit. These fits show small systematic deviations and yielded a slight inconsistency that was interpreted as stemming from the inapplicability of the exchange rate in the canonical coil state (shortest peptide) to the coil state seen for longer peptides capable of forming helices.

Thompson et al.¹⁹ constructed a kinetic zipper model for a similar alanine-based peptide (termed FS-peptide)²⁰ to simultaneously interpret data from laser T-jump experiments and CD. They found that the equilibrium data could be equally well reproduced by different parameter sets but that relaxation rates were only consistent with values of the T-independent nucleation parameter that are significantly larger than those reported by Rohl and colleagues.¹⁷ In their model, Thompson et al. were, however, restricted to the assumption of only a single helical segment being allowed to form. Examples, such as the three studies mentioned above, have led to the transferability of parameters derived from LR models being questioned.^{21,22}

Based on the extensive literature on the subject, several assumptions inherent to the application of LR models to helix–coil transition data emerge as questionable:

- (1) Even in the absence of helix formation, independence of the backbone angles of individual residues does not hold.^{22,23}

- (2) Helix stability does not just depend on hydrogen bonds but encompasses solvation and hydrophobic terms.^{24–26}
- (3) Scattering experiments and in silico studies have proposed that the single-sequence approximation is misleading even for relatively short peptides.^{27,28} It appears quite likely that helix bundles form through stabilization by tertiary interactions that are not representable in LR models. Such interactions are also one possible explanation for the observed length-dependent propagation behavior of α -helices.^{29–31}
- (4) As a corollary to the previous point, it is worth mentioning that LR models predict that very long helices are extremely stable, which contrasts with the low prevalence of long helices in biological systems: The likelihood of observing helices longer than 15 residues in globular proteins decreases rapidly,²⁸ and even putative coiled-coil domains rarely exceed 150 residues despite the presence of stabilizing and specific tertiary interactions.³² Of course, these data provide indirect evidence only as the impacts of evolutionary pressures vs physicochemical properties cannot be delineated.

In addition, applications of LR models to molecular simulation data have revealed that in almost all molecular force fields, the statistical likelihood of occupying the region of Ramachandran space compatible with α -helical hydrogen bonds is larger than proposed nucleation parameters suggest. Nucleation parameters are routinely overestimated when analyzing in silico data,^{5,33,34} and this constitutes either a fundamental error in force fields or a disconnect between in vitro and in silico interpretations of helix nucleation.

In this contribution, we employ molecular dynamics simulations in an all-atom representation of the FS-peptide (acetyl-A₅(AAARA)₃A-N-methylamide) coupled to the recently developed ABSINTH implicit solvation model.³⁴ Our aim was to generate a diverse but statistically sound set of data that highlight limits of applicability and interpretability of LR fits and parameters to computational and atomistic sampling of the temperature-dependent helix–coil transition. We employ a wide range of simulation temperatures and compare models differing in the imposed rigidity of backbone bond angles to explore the thermodynamics of the transition in richer detail. The known impact of such constraints³⁵ is found to be large and is affecting qualitative features of the sampled ensembles as well. Using our simulation data, we show that LR fits yield results that are unsatisfactory either in terms of parameter interpretability or in terms of fit accuracy. We highlight the lack of transferability by showing that the temperature dependence of the fitted helix propagation parameter cannot be connected easily to the bulk behavior of the peptide. Finally, we suggest additional tests and alternative routes for analyzing in silico data, the most important one being the inclusion of the mean number of isolated residues in the α -helical basin, $\langle N_1 \rangle$, in the LR fitting.

METHODS

Simulation Design. The FS-peptide (acetyl-A₅(AAARA)₃A-N-methylamide)²⁰ was enclosed in a spherical droplet of 40 Å radius along with explicit sodium and chloride ions compensating the peptide's positive charge and adding a background electrolyte concentration of ~150 mM. Starting configurations were random aside from satisfying excluded volume requirements. The effects of water were described by the ABSINTH implicit

solvation model,³⁴ which is a group transfer-based model similar in spirit to the EEF1 model³⁶ and based in parts on the OPLS-AA/L force field³⁷ (see Methods in Supporting Information for further details). The simulations integrated Langevin equations of motion at constant volume with a time step of 2.5 fs and a universal atomic friction coefficient of 1.2 ps^{-1} .³⁸ With these settings, integration was stable, and net temperature artifacts due to integrator, cutoff, and other noise terms assumed maximal values of $\sim 4\text{K}$ for the highest temperature (see below). The use of a Langevin integrator neglecting hydrodynamic interactions with artificially low friction in conjunction with an implicit solvent model means that the resultant conformational dynamics will not be physically realistic. The motivation for this setup lies in obtaining converged equilibrium data of the thermodynamics of the helix–coil transition as a function of temperature, which allow rigorous assessment of LR models.

To additionally enhance sampling, we employed the replica–exchange (RE) technique³⁹ and constructed two overlapping schedules each consisting of 16 temperatures. The low-temperature schedule used 220, 227, 234, 242, 250, 259, 268, 278, 288, 299, 310, 322, 334, 347, 360, and 374 K, and the high-temperature schedule used 260, 268, 276, 284, 292, 300, 310, 320, 330, 340, 350, 360, 375, 390, 410, and 440 K. Exchanges between neighbors were attempted every 25 ps in either case. The average acceptance probability for the swap moves generally exceeded 33% except for terminal replicas. The low exchange attempt frequency was intended, and the results show that sampling is robust regardless. Comparison of results from the two completely independent runs across the overlapping region allows a simple and rigorous assessment of sampling quality. It should be noted that implicit solvents do not exhibit phase transitions, thereby allowing the use of unusual temperatures. An exact mapping of simulation temperatures to realistic ones is typically not possible. Specifically for the ABSINTH continuum solvation model, temperatures between 280 and 350 K may be reasonably well-represented,³⁴ but the primary reason for using “unphysical” temperatures lies in our aims to create as diverse an ensemble of helical and coil states as possible and to optimize benefits from RE sampling to obtain statistically sound data.

Residue-based neighbor lists were recalculated every 5 steps, and interactions were generally truncated at 12 Å. Interactions between residues carrying a net charge were not truncated at all but instead computed in a monopole approximation if their distance exceeded 12 Å. The total simulation length of an individual temperature replica was always 250 ns, with the first 50 ns being discarded as equilibration. Two different sets of holonomic constraints were enforced during integration (see below). All simulations were performed using the homegrown CAMPARI software package.⁴⁰ The data for alanine dipeptide in Figure 7 were extracted from simulations of 125 ns in length. With the exception of the absence of any ions, these runs used identical conditions and settings and were performed independently for either set of constraints.

Constraints. We simulated the FS-peptide using two different sets of holonomic constraints enforced during integration of the equations of motion via the SHAKE algorithm.⁴¹ The first set constrained the lengths of all covalent bonds. This corresponds to a standard setup in molecular dynamics applications. The second set specifically rigidified backbone bond angles by introducing additional distance constraints between C_α and O, C_α and HN, N and C_β , C and C_β , N and C, C_{i-1} and C_{i+1} , and C_α and N_{i+1} . Even though the coupling between constraints is

increased, this set is still comfortably solvable by SHAKE. We used a relative tolerance of 10^{-4} and verified that the corresponding internal degrees of freedom were in fact constant throughout the simulations.

It should be noted that we ignored contributions to the equilibrium populations stemming from the mass-metric tensor determinant.⁴² Given that fixed bond lengths are not typically considered as a source of bias error and that we introduce only a subset of possible bond angle constraints, we assume that the combination of a stochastic dynamics integrator and a structured energy landscape renders potential artifacts minor.^{43,44} Support for this assumption is presented in Figure S1 in the Supporting Information, where we compare molecular dynamics to Monte Carlo data. The latter is based on an implementation³⁴ that rigidifies all bond angles (and some dihedral angles) and is inherently free of mass-metric tensor artifacts due to the absence of momenta. For the polypeptide backbone, it is therefore very similar to the case with rigidified backbone angles shown here. Consequently, quantitative similarity is expected and largely seen in Figure S1, Supporting Information. Formulations incorporating explicit corrections for mass-metric tensor artifacts exist but require dedicated integrators.^{44,45}

Analysis of Simulation Data. Statistics for all data were collected every 25 ps. The α -helical region of Ramachandran space was defined identically to previous work.³⁴ Define secondary structure of proteins (DSSP) statistics were collected by assigning secondary structure based on hydrogen-bond patterns using the actual trajectory coordinates for amide hydrogen atoms. The default cutoff criteria employed by Kabsch and Sander⁴⁶ were used throughout, but numerical tests (not shown) revealed the sensitivity of altering the energetic cutoff for hydrogen bonds from -0.5 to -0.3 and -1.0 kcal/mol, respectively, to be insignificant compared to the differences between flexible and rigidified models or between measures of helicity in Figures 1 or S1, Supporting Information. Helical segments in DSSP require at least two, consecutive hydrogen bonds of $i \rightarrow i + 4$ registry. This means that three-residue segments are missed in the DSSP analysis, which in LR theory are assumed to possess one helical hydrogen bond. Furthermore, one- and two-residue segments are not accounted for at all. Both methods can theoretically yield false positives and false negatives, and this is partially intended: Torsional statistics are purely based on inference, and any three-residue segment assigned as helix may easily be in a conformation not amenable to hydrogen-bond formation. Conversely, α -helical hydrogen bonds may be formed even when not all three intervening residues are in the torsional basins due to compensatory effects. DSSP assignments on the other hand imply that not all hydrogen bonds throughout a helix may satisfy the significance cutoff, but that the residues are treated as a single helix nonetheless. Conversely, two consecutive hydrogen bonds may both be barely within the cutoff and may not correspond to a proper α -helical segment.

Length-dependent statistics for helical segments (continuous residues in helical conformation as determined by either DSSP or torsional occupancy) were collected and used to determine $\langle N_s \rangle$ and $\langle N_1 \rangle$. For each encountered segment, the contribution to $\langle N_h \rangle$ was inferred as $l_s - 2$, where l_s is the length of the corresponding segment. To be able to use DSSP-based statistics consistently, counts for three-residue segments contributing to $\langle N_h \rangle$, for two- and three-residue segments contributing to $\langle N_s \rangle$, and for one-residue segments constituting $\langle N_1 \rangle$ were taken from the torsional assignment instead (this gives rise to the “DSSP corr.”

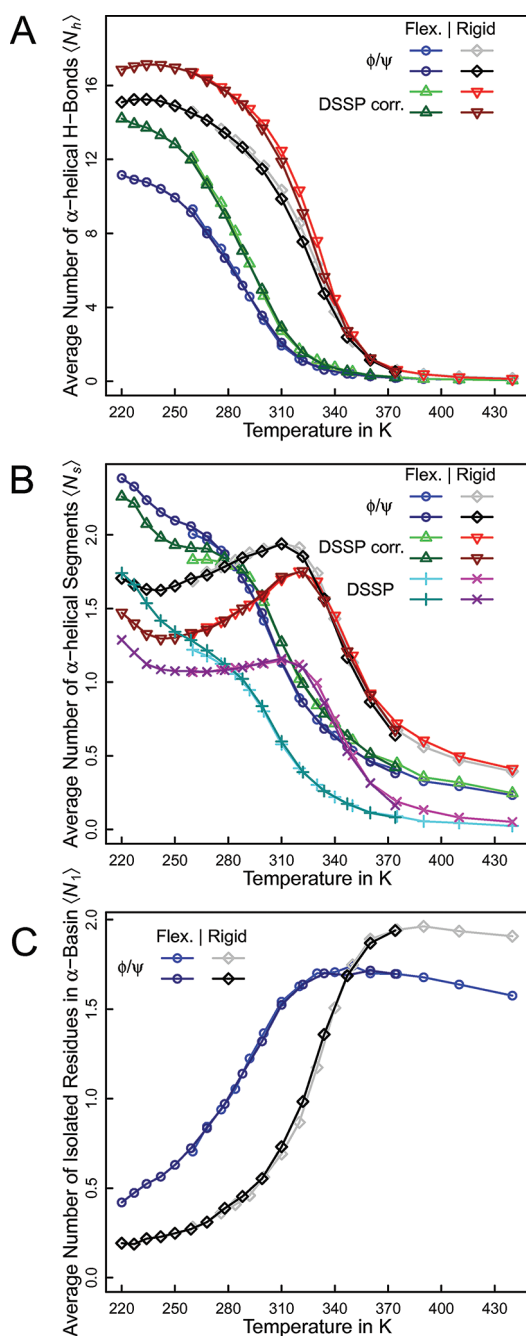


Figure 1. Quantification of helical content as a function of temperature for the FS-peptide using two different sets of holonomic constraints during the simulations. Panel A shows the mean number of α -helical hydrogen bonds, $\langle N_h \rangle$, inferred from either torsional statistics (“ ϕ/ψ ”) or DSSP assignments for the FS-peptide with corrections for three-residue segments (see Methods Section). Data for either set of constraints are indicated in the figure legend as “Rigid” (backbone bond angle constraints) and “Flex.” (no bond angle constraints). Panel B plots the mean number of distinct segments with at least two consecutive residues in helical conformation, $\langle N_s \rangle$. DSSP data not including the corrections are shown in addition to the rest. Panel C shows the average number of single residues in helical conformation surrounded by residues in nonhelical conformation, $\langle N_1 \rangle$. By construction, only data based on torsional statistics are available. In all plots, darker colors correspond to the replica exchange molecular dynamics (REMD) run across a lower set of temperatures and lighter colors to the higher temperature run. Lines are drawn as a guide to the eye only.

data set in Figures 1, 3, and 6 and S1, S3, and S4, Supporting Information). We believe it is important to include two-residue segments in the counting in contrast to suggestions in the recent literature.³ This is because otherwise $\langle N_h \rangle$ and $\langle N_s \rangle$ become more closely correlated, and less information than possible is being utilized.

Fitting Procedure. In all fits, the chosen model was fit to the data by a Monte Carlo procedure that allowed randomization over a reasonable interval (10% likelihood, 50% for the parameter f_3 that we introduce in eq 9 below) or stepwise perturbations (90% likelihood, 50% for f_3) of the fit parameters. All parameters were fit simultaneously, and a new set of values was accepted whenever the metric of goodness of fit was improved. The latter was defined as the normalized root-mean-square (rms) deviations of the two or three fitted quantities, viz., either $\langle N_h \rangle$ and $\langle N_s \rangle$ or $\langle N_h \rangle$, $\langle N_s \rangle$, and $\langle N_1 \rangle$. The normalization values were 19, 2, and 2, for $\langle N_h \rangle$, $\langle N_s \rangle$, and $\langle N_1 \rangle$, respectively. Normalized rms deviations were chosen to achieve a balanced impact of all three quantities irrespective of their value. The fits were generally highly reproducible and did not depend on the initial guess, indicating that a unique optimal solution given the metric of goodness exists. If this was not the case, it is noted in the text.

RESULTS AND DISCUSSION

In published computational work, connections to LR theory are usually made by parsing segment distributions for the peptide in question with respect to the α -basin which is defined by some heuristic.^{3,5,33} From this, $\langle N_s \rangle$ and $\langle N_h \rangle$ are estimated by assuming that, just like the LR stipulation, three consecutive residues in helix conformation will yield a hydrogen bond. This is an indirect estimation, and we show in Figure 1 how such inference compares to more direct estimates based on DSSP hydrogen-bond energies.

Cooperative Helix Melting and the Influence of Rigid Backbone Bond Angles. Figure 1A shows results from two independent temperature RE runs each for two different sets of constraints using both DSSP and torsional estimates of the number of α -helical hydrogen bonds. The first noteworthy point is the excellent congruency between the two independent RE runs. Since this constitutes convincing evidence toward the statistical reliability of the data, error estimates from block averaging, which would inherently be less rigorous indicators, are omitted for reasons of clarity from this and all further plots.

What impact does backbone rigidity have on the helix–coil transition? For both sets of constraints, the peptide shows a well-defined melting transition with increasing temperature. The loss of α -helical hydrogen bonds appears cooperative in either case, but, as is observed experimentally,^{19,47} occurs over a relatively broad temperature range. If bond angles along the backbone are rigidified, both the melting temperature and the limiting helical content in the helix phase experience a substantial upshift. This is true irrespective of whether hydrogen bonds are inferred by the DSSP algorithm or based on torsional segment statistics. The DSSP-based values are generally larger. This leaves at least two possibilities that are mutually compatible: On average the inference from torsional statistics misses hydrogen bonds (false negatives) and/or the DSSP inference overestimates numbers of hydrogen bonds (false positives, see Methods Section for details).

Panel B shows that there are qualitative differences between the two ensembles as well. With only bond lengths constrained,

the average number of segments with at least two consecutive residues in helical conformation increases continuously with decreasing temperature. This suggests that the high flexibility of the chain favors conformations containing two or more shorter helices. Conversely, the introduction of angle constraints in the polypeptide backbone appears to stabilize conformations with just a single helix over a wider range of temperatures leading to an actual decrease in the number of helical segments when reducing the temperature from 300 to 250 K. The uncorrected DSSP-derived segment statistics are not applicable to LR theory since conformations with exactly two or exactly three residues in helical conformation are missed due to the lack of the two hydrogen bonds required according to DSSP (see Methods Section). They can, however, be used to quantify the actual number of well-defined helical segments. This confirms that the qualitative dissimilarity between the two ensembles is robust. If we add the counts from torsional statistics for those two- and three-residue segments to the DSSP counts (“DSSP corr.” in the legend), the data for the case of flexible backbone angles are mutually consistent, i.e., the omissions of shorter segments implied by the DSSP algorithm are able to approximately explain the discrepancy in the data. This is *not* true for the case of rigidified backbone angles. Here, the discrepancy seems to stem mostly from a single, long helix being mistakenly broken into two or more pieces by the inaccuracy of the inference of hydrogen bonds based on torsional segment statistics. Later, we will therefore use both data sets.

Lastly, Figure 1C shows the number of single residues that are in helical conformation with both neighbors not being in helical conformation (“one-residue segments”). This is a complementary readout to the data in panel B and can, just like the other two, be directly estimated using the LR formalism:

$$\langle N_1 \rangle = \frac{\partial \ln Z}{\partial \ln \nu_{32}} \quad (5)$$

Note that ν_{32} is the (only) element of the matrix corresponding to three-residue sequences “chc” (coil, helix, coil), i.e., an isolated, helical residue. We will use this readout, which we are unaware of having been employed in the recent literature, and its characteristic temperature dependence below as a weakly dependent test for fits obtained using eqs 4.

In summary, the data in Figure 1 show that bond angle constraints have a profound impact on the nature of helix-rich ensembles even though the melting transition itself may be robust. It is worth pointing out, however, that the differences observed here are still smaller than those seen when comparing different force fields to one another^{3,5,48} or when comparing explicit to implicit solvent data.³³ It may be argued that increased local flexibility leads to access to larger parts of the Ramachandran map. Inspection of the corresponding data for alanine dipeptide (not shown) supports this statement and allows the tentative hypothesis that increased likelihood of helix nucleation leads to the shift in the melting transition upon rigidification of backbone bond angles. Importantly, the increased flexibility could also influence segment statistics in an artificial manner given the use of the same definition of the α -basin in either case. This is where the complementary DSSP analysis is important that shows consistent differences between flexible and rigid cases but should not be affected in a similarly straightforward manner by increased local flexibility. In fact, sensitivity analyses (not shown) emphasize the robustness of DSSP estimates with respect to changes in cutoff criteria.

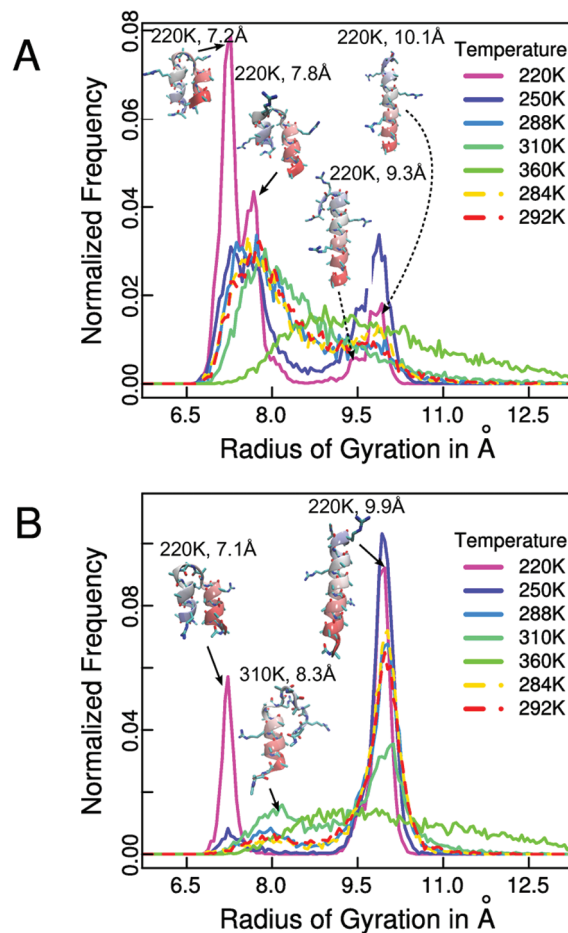


Figure 2. Histograms of radii of gyration (R_g) of the FS-peptide at different temperatures. Panel A shows the data for the case without any bond angle constraints, and panel B for the case with backbone bond angle constraints. The bin size for the construction of the histograms was 0.05 Å. All data drawn with solid lines are extracted from the low-temperature REMD run. To illustrate the statistical reliability of the data, we plot as dashed lines the two temperatures closest to 288 K found in the high-temperature REMD schedule (284 and 292 K). Clearly, the differences between substantially different temperatures vastly exceed the level of statistical noise in the data. To illustrate some of the dominant peaks, cartoon representations of individual structures along with their parent temperature and actual radius of gyration are given. These graphics were generated using VMD.⁶³

As a corollary, we do not believe that it is possible to tune the cutoff parameters for the two analyses types to make the resultant estimates of helicity mutually consistent. In fact, qualitative differences should persist on account of the fundamentally different information utilized and DSSP’s built-in fault tolerance. In this context, it should be stressed that the definition of the statistical weight w in LR theory does not require a specific interpretation in terms of dihedral angles that matches the one implied in the definition of v .

Single Helix or Collapsed Bundles? What is the nature of the qualitative differences observed between the two helix-rich ensembles? The data in Figure 1 suggest that with increased backbone flexibility, the peptide is more likely to form collapsed bundles of multiple helical segments, whereas the single helix is the dominant state with rigidified backbone angles. Figure 2 shows distributions of the radii of gyration for either case at a few

different temperatures. In the coil regime (360 K), comparison of panels A and B, shows that the two distributions are broad and very similar indicating that extended and disordered structures are populated in either case. In the helical regime (≤ 288 K), however, substantial differences are found. The distributions are generally multimodal with the peak at about 9.8 Å corresponding to the single, extended α -helix and the sharp peak at ~ 7.2 Å corresponding to the symmetric two-helix bundle (“helix–turn–helix”). These states and their sizes are perfectly consistent with the work of Zhang et al.,²⁸ who report 10.2 Å for the straight helix and 7.2 Å for a helix–turn–helix conformer (symmetric bundle) based on implicit solvent molecular dynamics simulations using an AMBER force field.

In the presence of just bond length constraints (panel A), the straight helix is never populated in dominant fashion, and bundles are more prominent. Its population appears to increase with temperature before melting occurs (above 300 K) presumably on account of the lessened drive to collapse. Conversely, with a rigidified backbone, the dominant helical state is the single helix. Here, the probability of observing partially collapsed states with radii of gyration of 7–8 Å seems to increase with increasing temperature when compared to the data at 250 K. If the temperature is dropped even further, a secondary transition sets in, in which the single helix collapses to form the two-helix bundle. This transition is also apparent in panels A and B of Figure 1. Complex coupling of coil-to-globule and helix–coil transitions has been observed for simplified models.^{49–51} One may ask whether the artificially low temperatures coupled with explicit representation of counterions influence these results, but an analysis of both ion–ion and peptide–ion pair correlation functions indicates that ions remain largely inert with very little direct binding at all temperatures (see Figure S2, Supporting Information).

LR Fitting. Next, we show that it is possible to fit a LR model to the data for just $\langle N_s \rangle$ and $\langle N_h \rangle$ by using eq 4 if no limits are placed on the values ν and w can assume. In Figure 3A and B, we show most of the same data as in Figure 1 as solid lines along with the fitted values (symbols). There are two fits, one to the data obtained from torsional inferences and the other to the data obtained from DSSP inference. Obviously the quality of the fit is arbitrarily good in either case suggesting that the two LR parameters are sufficiently independent. However, panel C shows that the resultant values for the LR parameters are inconsistent with the observed propensity to form isolated residues in helical conformation (see eq 5). $\langle N_i \rangle$ appears to be consistently overestimated when using the fitted values for ν and w , more so for the torsional case than for the DSSP estimates. This indicates that the obtained nucleation parameters are generally too large.

In panels A and B of Figure 4, we plot the actual values for ν and w resulting from the aforementioned fitting, respectively. We find the conjecture that large nucleation parameters cause an overestimation of $\langle N_i \rangle$ to be qualitatively confirmed. The nucleation parameter traces the temperature dependence of the propagation parameter irrespective of the constraint set employed or the data set fit to. At low temperatures, it assumes values that are indeed nonsensically large if one considers that the nucleation parameter should be related to the likelihood of visiting the α -region of ϕ/ψ -space in the absence of any hydrogen bonds. Conversely, in the coil region, the assumed values appear reasonable and close to the estimation of Thompson et al.¹⁹ of ~ 0.127 (page 9208, $\sigma_{zB} \sim 0.01$, and $\sigma_{zB} = \nu^2/(\nu + 1)^4$). Interestingly,

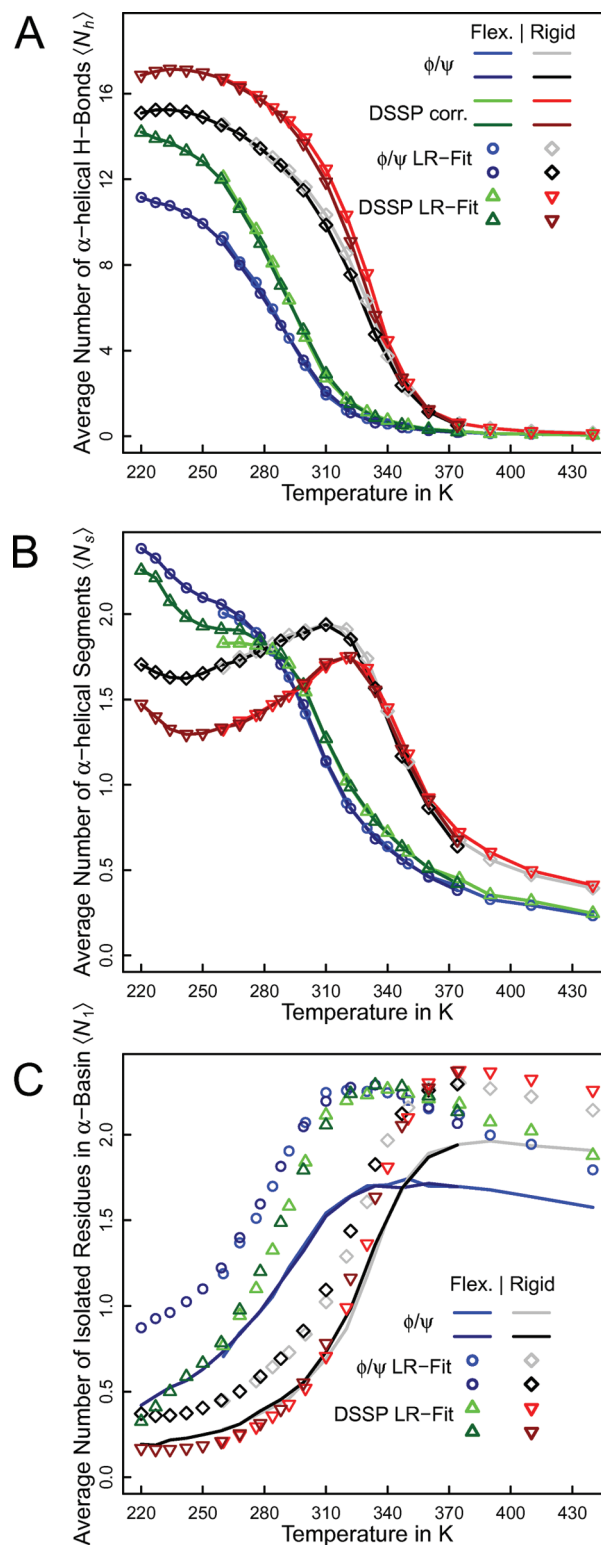


Figure 3. Quality of fits of LR theory to helical content data as a function of temperature for the FS-peptide using two different sets of holonomic constraints. The LR parameters obtained by these fits are plotted in Figure 4. To illustrate goodness of fits, solid lines are identical to those in Figure 1 and show the data measured directly from the simulations. Conversely, best-fitted values resulting from imposing the LR model are shown as symbols only (fits performed using eqs 4). Panel B uses the same legend as panel A.

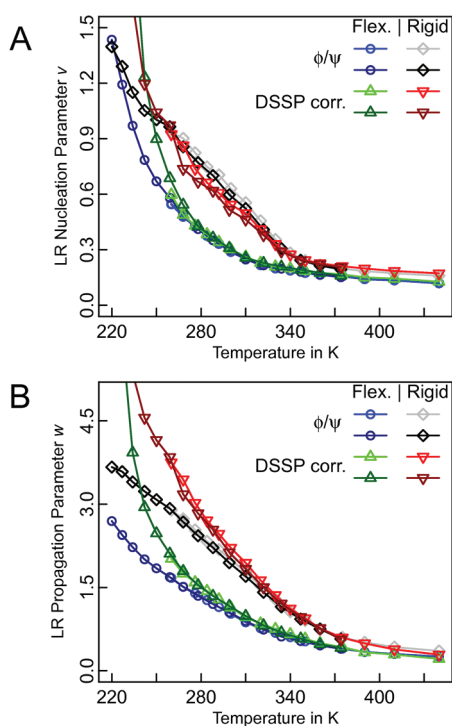


Figure 4. LR nucleation (A) and propagation (B) parameters as a function of temperature. Two types of fits are shown that use either DSSP-derived or torsional values for $\langle N_h \rangle$ and $\langle N_s \rangle$. The values shown give rise to the predictions shown as symbols in Figure 3 by using eq 4. The set of constraints enforced and the data set used are indicated in the legend similar to Figure 1. Note that low-temperature data for DSSP-based fits are cut off to allow visualization of all data in the same plot. They both continue to increase monotonously when further reducing the temperature.

given a set of constraints, the nucleation parameter seems to be mostly independent of the data set used (“DSSP corr.” vs “ ϕ/ψ ”) all the way down to temperatures of ~ 265 K, even though divergence of the fitted quantities occurs already at much higher temperatures in Figure 3A and B. Conversely, in Figure 4B, it is worth noting that the values of w and its dependency on temperature do appear to depend significantly on the data set used for fitting, suggesting that any enthalpy estimates using $\ln w$ will lack robustness (see below).

One may ask whether it is possible to fit to all three quantities ($\langle N_h \rangle$, $\langle N_s \rangle$, and $\langle N_1 \rangle$) with only one or two free variables. Figure S3, Supporting Information, shows that, when assuming a constant value for the nucleation parameter of 0.127, the quality of the fit drastically deteriorates. Essentially, it is impossible to predict correctly the values for $\langle N_s \rangle$ if only w is allowed to vary.³ Even though the agreement for $\langle N_1 \rangle$ may be improved due to inclusion in the fitting procedure, it is overall very clear that LR predictions are unable to explain the data. As suggested by Figure 2, the largest discrepancy arises on account of the inability to represent the stabilization of helical bundles ($\langle N_s \rangle$ significantly larger than unity). An unconstrained fit in v/w -space masks this inapplicability by producing large values for v . This is almost certainly the reason why *in silico* data that match melting temperature and overall helicity well universally exhibit large values for v when analyzed with LR theory.^{5,33,34} This inapplicability is masked of course if only data are analyzed that correspond to the transition and coil regimes but not to temperatures significantly

below the observed melting temperature, or if the force field does in fact produce strictly LR-like results.^{3,48}

Figure S4, Supporting Information, shows that the overall fit, as seen in Figure 3, can be improved when including $\langle N_1 \rangle$ as a fitted quantity. However, this may lead to a deterioration of fitting quality specifically for $\langle N_h \rangle$. Interestingly, both types of fits for either system now tend to agree more with the DSSP-derived hydrogen-bond counts. This is despite the fact that the values for $\langle N_1 \rangle$ are derived exclusively from torsional occupancies and indicates that the DSSP-derived statistics, which include torsional data for short segments (see Methods Section), may intrinsically be more consistent on account of the fact that they are much less prone to assign false breaks within long helices. In fact, overall fit quality is fairly good for the two DSSP-based fits. However, the values for the nucleation parameters remain large and exhibit even stronger dependencies on temperature. There are two ways to compensate the overestimation of single residues in α -conformation seen in Figure 3: making helices very stable (w large) or making the nucleation parameter so large that it is more likely to see two or more consecutive residues in helical conformation rather than one purely on account of v . Both paths are explored in Figure S4, Supporting Information; the former for DSSP and the latter for torsional statistics. This is an exacerbated demonstration of blind fitting yielding parameters that are impossible to interpret physically.

van't Hoff Analysis. The enthalpy change associated with the formation of a single hydrogen bond, ΔH_{hb} , is one of the parameters used most often to characterize the helix–coil transition experimentally. It is accessible from calorimetric experiments, and most recent estimates for alanine and alanine-like residues report a value of -0.9 kcal/mol⁵² with earlier values being slightly larger (-1.3 kcal/mol).⁵³ For experiments that directly measure helix content (e.g., CD), it is common to extract ΔH_{hb} from a van't Hoff plot by assuming the following temperature dependence for $\ln w$:^{8,10,11,48,54}

$$-\beta \cdot \Delta G_{hb} = -\beta \cdot \Delta H_{hb} + \Delta S_{hb}/R = \ln k \approx \ln w \quad (6)$$

Here, the subscript “hb” indicates that the process is interpreted to correspond to the addition of a single, α -helical hydrogen bond. We next critique the interpretation of $\ln w$ in eq 6 to arrive at a conclusion relevant to all LR-based analyses of helix–coil transition data.

If we consider a Schellman model⁵⁵ by assuming that only a single continuous helix is formed at any time and that no other residues, on average, reside in the helical basin at all, then the two-state equilibrium constant for the equilibrium between all-coil and all-helix states can be constructed as a product of stepwise constants:

$$K_{ch}^{cum} = \prod_{i=0}^{N_r-1} k_i = N_r v \cdot \frac{N_r-1}{N_r} v \cdot \frac{N_r-2}{N_r-1} w \cdot \dots = v^2 w^{N_r-2}$$

with $k_i = \frac{N_r-i}{N_r-i+1} \cdot w$ and $K_i = (N_r-i) \cdot v^2 w^{i-1}$ for $i > 2$

(7)

The statistical weight of a given sequence of N_r residues flanked by peptide bonds is the product of its residue weights, where the weight factor of a residue in the coil state, u , is set to 1 (normalization). Hence, a sequence “hhhcc” has total weight $v^2 w$ and sequence “hhhhc” has weight $v^2 w^2$. The combinatorial factors are simply related to the number of unique sequences that can accommodate a helical stretch of a given length. General

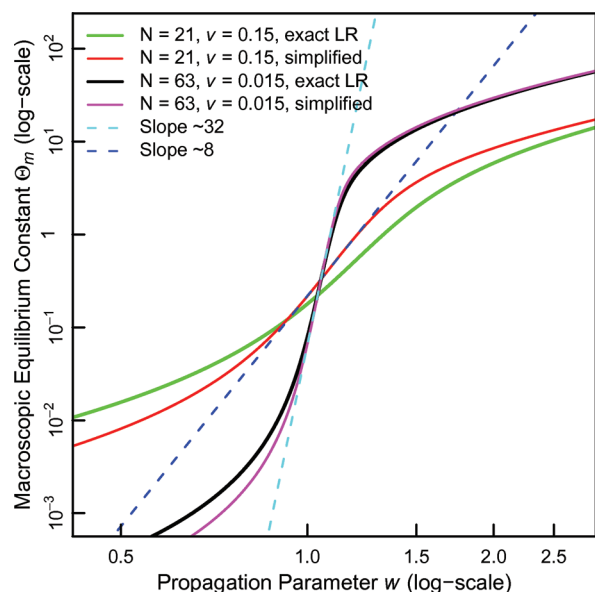


Figure 5. Comparison of the simplified single-sequence model (eq 7) to the exact LR formalism for two model systems. We show plots of Θ_m (see text) as a function of w when estimated using either the exact formula (eq 4) or the simplified version (eq 8). Agreement between the two depends on system parameters. Dashed lines indicate linear fits to the regions of maximal slope observed for the simplified model. The derivative can also be obtained analytically (see Supporting Information). Numerical tests show that the maximal slope does not approach $N_r - 2$ even when ν is reduced by another 2 orders of magnitude for the case with $N_r = 63$.

forms for stepwise (k_i) and cumulative (K_i) equilibrium constants are provided in eq 7, the latter being referenced to the all-coil state. Clearly, the stepwise constants suggest the approximation in eq 6 to be applicable when considering an isolated growth step as long as the helix is nucleated and not yet close to its maximum length. The expected slope in a double logarithmic plot of K_{ch}^{cum} and w would indeed be $N_r - 2$ supporting the view that values obtained via eq 6 correspond to numbers per hydrogen bond. However, this equilibrium between the all-coil and all-helical states is monitored neither experimentally nor computationally; in both cases, ensemble averages are used to determine w . For $\langle N_h \rangle$, the simple model above yields

$$\langle N_h \rangle = \frac{\sum_{j=2}^{N_r-1} K_j \cdot (j-1)}{Q} \quad \text{with } K_n = \prod_{i=0}^n k_i \quad \text{and} \quad Q = 1 + \sum_{i=0}^{N_r-1} K_i \quad (8)$$

We can thus construct a generalized equilibrium constant, Θ_m , for the helix–coil transition as $f_h/(1-f_h)$, where $f_h = \langle N_h \rangle / (N_r - 2)$, i.e., the fractional helicity, and compare it in terms of its dependency on w to data extracted from exact application of LR theory (see eq 4). This is shown in Figure 5 for two cases: the first corresponds to a scenario where the single-sequence model above should be reasonably applicable (small ν , larger N_r). Indeed, predictions from exact LR theory and from the simplified model agree very well. However, the relationship between the logarithms of Θ_m and w is complex. If we fit a line to the region of maximal variation (corresponding to states ranging from low to intermediate helicity), the resultant slope is only ~ 32 , i.e.,

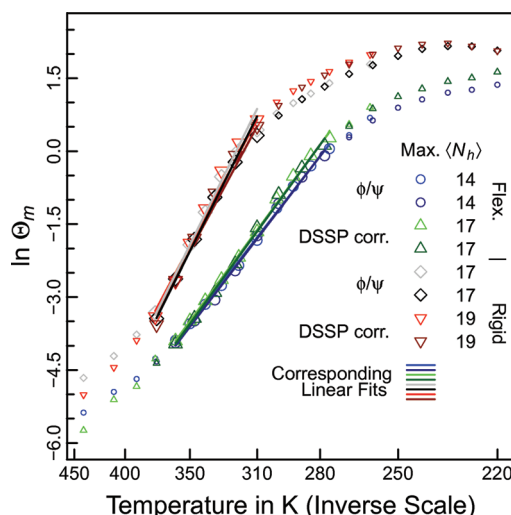


Figure 6. Van't Hoff determination of thermodynamic parameters of the helix–coil transition. Data are based on those in Figure 1A (see text). Linearity holds throughout the transition where both helix- and coil-rich states are populated to significant extent. With no angle constraints, the fitting region spanned from ~ 280 – 360 K, and with angle constraints we used ~ 310 – 375 K. These segments do in fact encompass the temperature regions exhibiting the largest change in Figure 1A. In the low-temperature region, secondary processes may prevent the van't Hoff assumption of temperature-independent enthalpy from being valid. The legend indicates the value assumed as the upper baseline for constructing f_h from $\langle N_h \rangle$ (see text). The obtained values are $\Delta S = -[34-35]$ cal \cdot mol $^{-1} \cdot$ K $^{-1}$ and $\Delta H = -[9.6-9.8]$ kcal \cdot mol $^{-1}$ for the case without bond angle constraints and $\Delta S = -[44-45]$ cal \cdot mol $^{-1} \cdot$ K $^{-1}$ and $\Delta H = -[13.8-14.5]$ kcal \cdot mol $^{-1}$ in the presence of angle constraints. The fits are no more dependent on the data set used than they are on the intrinsic accuracy of the data, which can be estimated by the differences obtained by independently fitting the low- and high-temperature REMD runs in each case. Lastly, values are not particularly sensitive to the definitions of upper baselines and temperature intervals. For example, the total variation is below 20% when including one additional temperature at each end or when changing the upper baseline from 14 to 19 for the case of flexible backbone and torsional data.

slightly more than half of the possible hydrogen bonds. It is therefore inaccurate to assume that application of eq 6 will yield values that can be interpreted as values per residue or per hydrogen bond (this would require a slope proportional to N_r). With parameters mimicking the system under study here, we find that the simple model becomes less applicable and that the slope for the full LR model is less than that found in the simplified model. Further numerical tests (see Figure S5, Supporting Information) clearly demonstrate that the maximum encountered slope has a nontrivial dependency on both N_r and ν , that it is always larger in the simplified model, and that it will often lie close to $(N_r - 2)/2$. This is an important point, as it means that results from fitting $\ln w$ in a van't Hoff-type plot should not be interpreted to be contributions per hydrogen bond.

For simulation data, we therefore advocate to construct van't Hoff plots directly from measured equilibrium constants as described above, where the problem of identifying baselines is negligible. In vitro, van't Hoff fits of $\ln w$ usually require the definition of baselines implicitly that can often be determined with better accuracy using cosolute titrations. In Figure 6, van't Hoff plots of the values of Θ_m constructed from the data for $\langle N_h \rangle$ in Figure 1A are shown over temperature regimes where linearity

holds. The lower baseline was always $\langle N_h \rangle = 0$, while the upper baselines we used are indicated in the legend. By this methodology, we obtain thermodynamic parameters for the entire process that are independent of whether DSSP or torsional statistics are used. The actual values agree well with literature estimates of $\Delta S = -36 \text{ cal} \cdot \text{mol}^{-1} \cdot \text{K}^{-1}$ and $\Delta H = -12 \text{ kcal} \cdot \text{mol}^{-1}$ ⁴⁷ and $\Delta S = -51 \text{ cal} \cdot \text{mol}^{-1} \cdot \text{K}^{-1}$ and $\Delta H = -14.8 \text{ kcal} \cdot \text{mol}^{-1}$ that are obtained in similar fashion directly from spectroscopic data.⁵⁶ The agreement is congruent with the fact that the estimated melting temperatures from experiment (290–306 K) overlap with the interval defined by the apparent melting temperatures of the two simulated ensembles (see Figure 1A). The total enthalpy gives rise to estimated values for ΔH_{hb} of -0.5 and $-0.75 \text{ kcal} \cdot \text{mol}^{-1}$ for flexible and rigidified backbones, respectively. These values are mutually consistent with the calorimetric estimate of $-0.9 \text{ kcal} \cdot \text{mol}^{-1}$ that by definition has to be larger in magnitude given that it will include contributions from factors not related to hydrogen bonding (most prominently overall peptide swelling). They are also consistent with the values obtained for fits to $\ln w$, which yield values between -1.0 and -1.3 kcal/mol experimentally,^{11,18} if we consider that ΔH_{hb} in such a case should really correspond to the enthalpy associated with the formation of *more than one* hydrogen bond (see above). Of course, the agreement between the particular computational model in use and experimental data at the level of thermodynamics may be fortuitous. It is noteworthy that the force field in use here implies discarding most of the dihedral angle potential parameters³⁴ that continue to be optimized elsewhere.^{3,5,57,58} Crucially, however, neither LR fits nor van't Hoff plots resolve potential discrepancies in mechanisms or dynamics of the helix–coil transition that could, for example, arise on account of the continuum solvation model lacking an appropriate description of water–peptide interfaces regarding wetting behavior, reorientation dynamics, etc.⁵⁹ It would therefore be ill-advised to arrive at conclusions on relative virtues of different computational models purely based on analyses like the ones presented here.

Modeling of Equilibrium between Single Helix and Multi-helix Bundles. Lastly, is there a simple way to improve the original LR model, which specifically addresses issues identified here? For conceptual illustration, we test here a nongeneralizable modification to the fitting procedure that leaves the LR framework intact at the expense of an additional parameter. We focus on the statistics derived from torsional segments only since DSSP statistics need to be augmented by data on short segments derived from ϕ/ψ -values.

Following some of the ideas in the work of Ghosh and Dill,⁶⁰ we consider the system to be in equilibrium between a three-helix “bundle” and a single helix. Then, we may approximately treat the three-helix bundle as three independent sequences of one-third the length of the original peptide:

$$\langle N_h \rangle = 3f_3 \cdot \frac{\partial \ln Z_{N_t=7}}{\partial \ln w} + (1-f_3) \cdot \frac{\partial \ln Z_{N_t=21}}{\partial \ln w} \quad (9)$$

The averages $\langle N_1 \rangle$ and $\langle N_s \rangle$ are computed analogously (see eqs 4). The new parameter f_3 is simply the fractional occupancy of the three-helix bundle and setting it to zero recovers the original fitting functions as used in Figures 3 and 4. How are nonzero values of f_3 interpretable? Essentially, we stipulate that there are reasons external to LR theory that “stabilize” helix interruptions. In a thermodynamic sense, these can be tertiary interactions stabilizing compact bundles. However, in a statistical sense, they can also be errors in the counting of helical segments

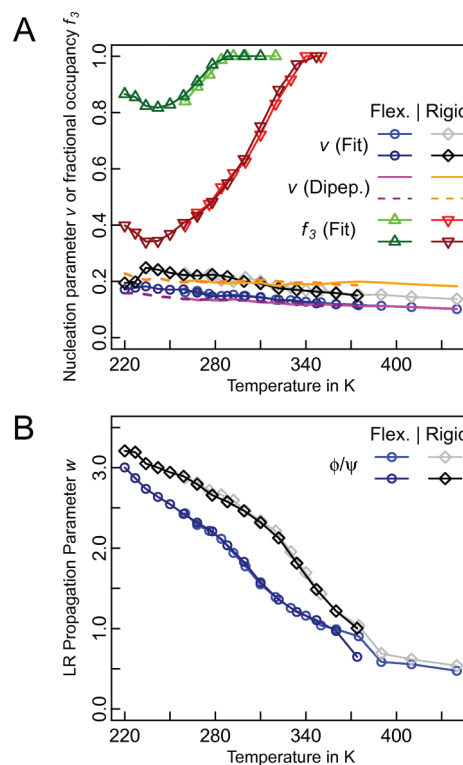


Figure 7. Fitted values v , w , and f_3 as a function of temperature when employing eq 9. Only fits to the data based on torsional inference are considered. Panel A shows the values for the nucleation parameter and f_3 in the same plot. Only those values for f_3 are shown that proved reproducible through multiple independent Monte Carlo fits (see main text and Methods Section). Along with the fitted values for v , we show the ratio of probabilities of occupying the helix and coil regions determined from simulations of alanine dipeptide, i.e., $p_h/p_c(T)$. The definition of the helical region was identical to the one used throughout to analyze data for the FS-peptide. Panel B shows the values for w . As in other figures, darker colors correspond to the low-temperature REMD run.

and their lengths. Evidence for both was presented above. Using eq 9 and treating f_3 as a free parameter, we obtain the fits and data in Figures S6, Supporting Information, and 7, respectively.

The first thing to note in Figure S6, Supporting Information, is that the overall fit quality is significantly improved over that shown in Figures S4, Supporting Information, which is of course expected due to the inclusion of an additional parameter. Nonetheless, f_3 is not able to explain all of the data consistently, as minor deviations are observed in the fitted values for $\langle N_h \rangle$ for the case with flexible bond angles. In Figure 7A, we show the obtained values for v and f_3 . Consistent with physical intuition, the nucleation parameter now assumes values in the interval from 0.1 to 0.25 and, for both systems, exhibits a very weak temperature dependence. This is despite the fact that no constraints were placed on the values v can assume during the fitting. We show that these values are reasonable for the computational model in use by explicitly computing the ratio of weights of the helical vs coil regions for alanine dipeptide as a function of temperature. As can be gleaned from Figure 7A, the agreement is profound. Both the sign of the temperature derivative and the differences between flexible and rigidified backbones are mirrored in the dipeptide data. We can also infer that differences in local backbone conformational properties may in fact be able to

explain the observed shift in the melting transition as was hypothesized above.

The values for f_3 are not particularly informative in the coil region since large differences have little impact on fit quality if long helical segments are generally unlikely to form. This means that the fits become ill-defined (not substantially dependent on f_3), and we omit those data points in Figure 7A. The model apparently suggests that the data are well-described by the three-helix bundle, in particular for the case of flexible bond angles. This is qualitatively consistent with Figure 2, in which the height of the peak at ~ 10 Å (single helix) strongly depends on the constraint set in use. The temperature dependence at low temperatures is consistent with Figure 2 as well in that bundled conformations are least likely at an intermediate temperature within the strongly helical region. In that sense, f_3 is physically interpretable. However, we wish to remind the reader that these fits are to quantities inferred from torsional statistics that are inherently prone to produce false negatives (see Methods Section and above). This may help to explain why in general the values for f_3 are large. Fitting this parameter may therefore simply represent a way to silently correct such faulty assignments. Unfortunately, the two effects are not easily deconvoluted. Along those lines, it may be interesting to ask whether a generalization of the model in eq 9 to arbitrary subsegment length distributions could produce even better results. The problem here is the limited data available for fitting a larger number of parameters. Figure S7, Supporting Information, shows a variant of eq 9, that can be fit unambiguously, producing inferior results. Lastly, it may be tempting to try to transform the data in Figure 2 into a direct and independent estimate for f_3 or related parameters, but such an effort would require the definition of a fair number of ad hoc structural criteria for clustering data.

CONCLUSIONS

This contribution makes a number of points that can be grouped into two categories. The first four all deal with the application of LR models to molecular simulation data and also with comparisons between in silico and in vitro results. Conclusions are as follows:

- (1) Estimates of the LR nucleation and propagation parameters are not directly comparable to those extracted from experimental data if the processes for obtaining those are different (Figures 3 and 4 and S3 and S4, Supporting Information). For example, it is invalid to perform an unconstrained fit to $\langle N_h \rangle$ and $\langle N_s \rangle$, as in Figures 3 and 4 for a single chain length, and compare it to estimates such as those by Rohl and Baldwin¹⁸ or Thompson et al.¹⁹ that use a fundamentally different construct of assumptions. Moreover, values for v and w that agree with experiment at a specific temperature may mask inaccuracies, and we recommend reporting melting temperatures and van't Hoff enthalpies instead (Figures 1 and 6).
- (2) Two checks are recommended: (i) mutual consistency of eligible helix–coil descriptors ($\langle N_h \rangle$ and $\langle N_s \rangle$) between torsional and DSSP inference and (ii) use of $\langle N_1 \rangle$ as either a weakly dependent test or an additional quantity to fit to (Figures 3 and S3, S4, and S6, Supporting Information). The robustness of estimation in particular of $\langle N_s \rangle$ will depend on the nature of the force field, and smaller deviations than those reported here may be found if the polypeptide backbone exhibits a larger amount of preorganization.³⁴

- (3) We show that it is misleading to interpret data from van't Hoff fits of $\ln w$ as quantities per residue or per hydrogen bond (Figures 5 and S5, Supporting Information). Of course, for similar procedures and identical systems, values obtained in such a way are still comparable to one another, but their physical meaning is not immediately obvious to us. In contrast, direct van't Hoff analyses of a generalized equilibrium constant, such as Θ_m , yield robust results that in this case also agree well with both IR and calorimetric estimates (Figure 6).^{25,47,52,56}
- (4) Lastly, we demonstrate that simple models can be found that preserve physical interpretability of fitted helix–coil parameters (Figure 7). It would be desirable to have a generalized framework for analyzing in silico data that satisfies the criteria spelled out above. One approach could be the ascending levels model of Lucas et al.⁵⁴ The problem thus far is that it is not routinely feasible to simulate reversible helix formation for many different peptides of differing lengths under a wide variety of conditions. Consequently, inconsistencies in the analysis are easily masked, and conclusions may be misleading.

The fifth and last point is more technical in nature:

- (5) Bond angle constraints alter the free energy landscape substantially and give rise to quantitatively and qualitatively different ensembles (Figures 1 and 2). As noted,³⁵ force field reparametrization will often be required to add (or release) such constraints. Therefore, they should not be viewed as independent entities controlling computational efficiency only.⁶¹ In contrast to backbone bond angle constraints, we did not observe strong changes of the kind seen in Figures 1 and 2 upon introduction of just bond length constraints (data not shown).

In summary, we suggest guidelines and checks for applying LR or similar theories to data obtained from atomistic simulations of helix-forming polypeptides. Ultimately, LR models may well be inapplicable to such data, and there is a clear need for a unified framework.^{60,62} We also believe that this work helps to reconcile some of the discrepancies in interpreting helix–coil transition data using the LR or similar formalisms, for example, when comparing in vitro to in silico data and also when comparing different sets of in vitro data to each other.

ASSOCIATED CONTENT

S Supporting Information. Illustration of the LR model using a simple example. Methods describing the force field and solvation model in more detail. Methods and plots (S1) on control simulations using Monte Carlo sampling. Additional plots showing ion pair correlation functions (S2), and LR fits and fitted parameters using different assumptions (S3–S4). Analytical derivations and numerical exploration of the model defined by eqs 7 and 8 (S5). Details on the interpretation of the partition function underlying eq 9. Quality of fits associated with Figure 7 (S6) and exploration of a related model (S7). This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: a.vitalis@bioc.uzh.ch. Telephone: +41446355597.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

We thank Dr. Rohit V. Pappu for critical comments regarding this work. This work was supported in part by a grant from the Swiss National Science Foundation to A.C. and a personal grant by the UZH Forschungskredit to A.V. Most of the simulations were carried out on the Schrödinger supercomputer administered by the IT services of the University of Zurich.

REFERENCES

- (1) Scholtz, J. M.; Baldwin, R. L. *Annu. Rev. Biophys. Biomol. Struct.* **1992**, *21*, 95–118.
- (2) Makhatadze, G. I. *Adv. Protein Chem.* **2006**, *72*, 199–226.
- (3) Best, R. B.; Hummer, G. *J. Phys. Chem. B* **2009**, *113*, 9004–9015.
- (4) Song, K.; Stewart, J. M.; Fesinmeyer, R. M.; Andersen, N. H.; Simmerling, C. *Biopolymers* **2008**, *89*, 747–760.
- (5) Sorin, E. J.; Pande, V. S. *Biophys. J.* **2005**, *88*, 2472–2493.
- (6) Garcia, A. E.; Sanbonmatsu, K. Y. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *99*, 2782–2787.
- (7) Ferrara, P.; Apostolakis, J.; Cafisch, A. *J. Phys. Chem. B* **2000**, *104*, 5000–5010.
- (8) Zimm, B. H.; Bragg, J. K. *J. Chem. Phys.* **1959**, *31*, 526–535.
- (9) Gibbs, J. H.; DiMarzio, E. A. *J. Chem. Phys.* **1959**, *30*, 271–282.
- (10) Lifson, S.; Roig, A. *J. Chem. Phys.* **1961**, *34*, 1963–1974.
- (11) Scholtz, J. M.; Hong, Q.; York, E. J.; Stewart, J. M.; Baldwin, R. L. *Biopolymers* **1991**, *31*, 1463–1470.
- (12) Bixon, M.; Scheraga, H. A.; Lifson, S. *Biopolymers* **1963**, *1*, 419–423.
- (13) Bixon, M.; Lifson, S. *Biopolymers* **1967**, *5*, 509–514.
- (14) Doig, A. J.; Chakrabarty, A.; Klingler, T. M.; Baldwin, R. L. *Biochemistry* **1994**, *33*, 3396–3403.
- (15) Shalongo, W.; Stellwagen, E. *Protein Sci.* **1995**, *4*, 1161–1166.
- (16) Kemp, D. S. *Helv. Chim. Acta* **2002**, *85*, 4392–4423.
- (17) Rohl, C. A.; Scholtz, J. M.; York, E. J.; Stewart, J. M.; Baldwin, R. L. *Biochemistry* **1992**, *31*, 1263–1269.
- (18) Rohl, C. A.; Baldwin, R. L. *Biochemistry* **1997**, *36*, 8435–8442.
- (19) Thompson, P. A.; Eaton, W. A.; Hofrichter, J. *Biochemistry* **1997**, *36*, 9200–9210.
- (20) Lockhart, D. J.; Kim, P. S. *Science* **1992**, *257*, 947–951.
- (21) Bierzyński, A.; Pawłowski, K. *Acta. Biochim. Pol.* **1997**, *44*, 423–432.
- (22) Jacobs, D. J.; Wood, G. G. *Biopolymers* **2011**, *95*, 240–253.
- (23) Pappu, R. V.; Srinivasan, R.; Rose, G. D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 12565–12570.
- (24) Shalongo, W.; Dugad, L. B.; Stellwagen, E. *J. Am. Chem. Soc.* **1994**, *116*, 2500–2507.
- (25) Taylor, J. W.; Greenfield, N. J.; Wu, B.; Privalov, P. L. *J. Mol. Biol.* **1999**, *291*, 965–976.
- (26) Ghosh, T.; Garde, S.; Garcia, A. E. *Biophys. J.* **2003**, *85*, 3187–3193.
- (27) Zagrovic, B.; Jayachandran, G.; Millett, I. S.; Doniach, S.; Pande, V. S. *J. Mol. Biol.* **2005**, *353*, 232–241.
- (28) Zhang, W.; Lei, H.; Chowdhury, S.; Duan, Y. *J. Phys. Chem. B* **2004**, *108*, 7479–7489.
- (29) Kennedy, R. J.; Tsang, K. W.; Kemp, D. S. *J. Am. Chem. Soc.* **2002**, *124*, 934–944.
- (30) Miller, J. S.; Kennedy, R. J.; Kemp, D. S. *J. Am. Chem. Soc.* **2002**, *124*, 945–962.
- (31) Wang, T.; Zhu, Y. J.; Getahun, Z.; Du, D. G.; Huang, C. Y.; DeGrado, W. F.; Gai, F. *J. Phys. Chem. B* **2004**, *108*, 15301–15310.
- (32) Rose, A.; Schraegle, S. J.; Stahlberg, E. J.; Meier, I. *BMC Evol. Biol.* **2005**, *5*, 66.
- (33) Nymeyer, H.; Garcia, A. E. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13934–13939.
- (34) Vitalis, A.; Pappu, R. V. *J. Comput. Chem.* **2009**, *30*, 673–699.
- (35) Chen, J.; Im, W.; Brooks, C. L., III. *J. Comput. Chem.* **2005**, *26*, 1565–1578.
- (36) Lazaridis, T.; Karplus, M. *Proteins: Struct., Funct., Bioinf.* **1999**, *35*, 133–152.
- (37) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.
- (38) Skeel, R. D.; Izaguirre, J. A. *Mol. Phys.* **2002**, *100*, 3885–3891.
- (39) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.
- (40) Vitalis, A.; Steffen, A.; Lyle, N.; Mao, A. H.; Pappu, R. V. CAMPARI, v1.0; SourceForge/Geeknet, Inc.: Mountain View, CA, 2010; <http://sourceforge.net/projects/campari> (accessed December 13, 2011).
- (41) Ryckaert, J. P.; Cicotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.
- (42) Fixman, M. *Proc. Natl. Acad. Sci. U.S.A.* **1974**, *71*, 3050–3053.
- (43) Perchak, D.; Skolnick, J.; Yaris, R. *Macromolecules* **1985**, *18*, 519–525.
- (44) Patriciu, A.; Chirikjian, G. S.; Pappu, R. V. *J. Chem. Phys.* **2004**, *121*, 12708–12720.
- (45) Mülders, T.; Swegat, W. *Mol. Phys.* **1998**, *94*, 395–399.
- (46) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577–2637.
- (47) Williams, S.; Causgrove, T. P.; Gilmanshin, R.; Fang, K. S.; Callender, R. H.; Woodruff, W. H.; Dyer, R. B. *Biochemistry* **1996**, *35*, 691–697.
- (48) Gnanakaran, S.; Garcia, A. E. *Proteins: Struct., Funct., Bioinf.* **2005**, *59*, 773–782.
- (49) Sikorski, A.; Romiszowski, P. *Biopolymers* **2003**, *69*, 391–398.
- (50) Varshney, V.; Carri, G. A. *Phys. Rev. Lett.* **2005**, *95*, 168304.
- (51) Nowak, C.; Rostiashvili, V. G.; Vilgis, T. A. *J. Chem. Phys.* **2007**, *126*, 34902.
- (52) Lopez, M. M.; Chin, D. H.; Baldwin, R. L.; Makhatadze, G. I. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 1298–1302.
- (53) Scholtz, J. M.; Marqusee, S.; Baldwin, R. L.; York, E. J.; Stewart, J. M.; Santoro, M.; Bolen, D. W. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 2854–2858.
- (54) Lucas, A.; Huang, L.; Joshi, A.; Dill, K. A. *J. Am. Chem. Soc.* **2007**, *129*, 4272–4281.
- (55) Schellman, J. A. *J. Phys. Chem.* **1958**, *62*, 1485–1494.
- (56) Ianoul, A.; Mikhonin, A.; Lednev, I. K.; Asher, S. A. *J. Phys. Chem. A* **2002**, *106*, 3621–3624.
- (57) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 712–725.
- (58) Mackerell, A. D., Jr.; Feig, M.; Brooks, C. L., III. *J. Comput. Chem.* **2004**, *25*, 1400–1415.
- (59) Laage, D.; Stirnemann, G.; Sterpone, F.; Rey, R.; Hynes, J. T. *Annu. Rev. Phys. Chem.* **2011**, *62*, 395–416.
- (60) Ghosh, K.; Dill, K. A. *J. Am. Chem. Soc.* **2009**, *131*, 2306–2312.
- (61) Feenstra, K. A.; Hess, B.; Berendsen, H. J. C. *J. Comput. Chem.* **1999**, *20*, 786–798.
- (62) Jacobs, D. J.; Dallakyan, S.; Wood, G. G.; Heckathorne, A. *Phys. Rev. E* **2003**, *68*, 061109.
- (63) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33–38.

Correction to “SHARC – *Ab Initio* Molecular Dynamics with Surface Hopping in the Adiabatic Representation Including Arbitrary Couplings” [*J. Chem. Theory Comput.* 2011, 7, 1253–1258]

Martin Richter,[†] Philipp Marquetand,^{*,†,§} Jesús González-Vázquez,^{*,†} Ignacio Sola,[†] and Leticia González^{†,§}

[†]Departamento de Química Física I, Universidad Complutense, 28040 Madrid, Spain

[‡]Institut für Physikalische Chemie, Friedrich-Schiller-Universität Jena, Helmholtzweg 4, 07743 Jena, Germany

Journal of Chemical Theory and Computation 2011, 7, 1253–1258

In a recent paper,¹ we lined out a novel *ab initio* molecular dynamics (MD) method termed SHARC, which is able to treat arbitrary couplings in molecular systems including all degrees of freedom. The basis is Tully’s surface hopping scheme,² as it is for several other methods, see e.g. refs 3–8. Laser-induced couplings were treated in the surface hopping formalism for the first time by Thachuk et al.³ and later on by Jones et al.⁵ as well as Mitrić et al.⁶ However, to our knowledge, we have treated spin–orbit coupling and dipole couplings simultaneously for the first time in MD. In a first test case, we validated the ability of our new method to describe laser coupling only. To this aim, we used two displaced harmonic oscillators. The same model was employed before by Mitrić and co-workers using the so-called FISH method in ref 36 of our paper (here ref 6); the latter methodology is shown to be a very good approximation to interpret some laser control experiments.^{9–12} Unfortunately, we did not point out clearly in ref 1 that the parameters of the test system were taken from the aforementioned source, which we hereby do. In ref 1, we repeated the simulations on the model system and obtained the same results as Mitrić and co-workers, in agreement with exact quantum dynamics. This means that the FISH and SHARC methods yield identical results for this case. Yet, there may be significant differences in the performance of the SHARC method compared to the FISH method for certain problems. Both methods, although technically different, serve to pursue the same goal, namely, to unravel unknown photochemical processes and mechanisms in large molecules including all degrees of freedom.

AUTHOR INFORMATION

Corresponding Author

*E-mail: philipp.marquetand@univie.ac.at; jgv@tchiko.quim.ucm.es.

Present Addresses

[§]Institute of Theoretical Chemistry, University of Vienna, Währinger St. 17, 1090 Vienna, Austria

REFERENCES

- (1) Richter, M.; Marquetand, P.; González-Vázquez, J.; Sola, I.; González, L. *J. Chem. Theory Comput.* 2011, 7, 1253–1258.
- (2) Tully, J. C. *J. Chem. Phys.* 1990, 93, 1061–1071.
- (3) Thachuk, M.; Ivanov, M. Y.; Wardlaw, D. M. *J. Chem. Phys.* 1996, 105, 4094–4104.
- (4) Maiti, B.; Schatz, G. C.; Lendvay, G. *J. Phys. Chem. A* 2004, 108, 8772–8781.

- (5) Jones, G. A.; Acocella, A.; Zerbetto, F. *J. Phys. Chem. A* 2008, 112, 9650–9656.
- (6) Mitrić, R.; Petersen, J.; Bonačić-Koutecký, V. *Phys. Rev. A* 2009, 79, 053416.
- (7) Shenvi, N. *J. Chem. Phys.* 2009, 130, 124117.
- (8) Tavernelli, I.; Curchod, B. F. E.; Rothlisberger, U. *Phys. Rev. A* 2010, 81, 052508.
- (9) Petersen, J.; Mitrić, R.; Bonačić-Koutecký, V.; Wolf, J.; Roslund, J.; Rabitz, H. *Phys. Rev. Lett.* 2010, 105.
- (10) Mitrić, R.; Petersen, J.; Wohlgemuth, M.; Werner, U.; Bonačić-Koutecký, V. *Phys. Chem. Chem. Phys.* 2011, 13, 8690.
- (11) Lisinetskaya, P.; Mitrić, R. *Phys. Rev. A* 2011, 83.
- (12) Mitrić, R.; Petersen, J.; Wohlgemuth, M.; Werner, U.; Bonačić-Koutecký, V.; Wöste, L.; Jortner, J. *J. Phys. Chem. A* 2011, 115, 3755–3765.

Published: December 13, 2011