

Computational Studies of the Luciferase Light-Emitting Product: Oxyluciferin

Luís Pinto da Silva and Joaquim C.G. Esteves da Silva*

Centro de Investigação em Química (CIQ-UP), Departamento de Química e Bioquímica, Universidade do Porto, Campo Alegre 687, 4169-007 Porto, Portugal

ABSTRACT: Firefly luciferase is the most studied bioluminescence system, and its catalyzed reactions have been relatively well characterized. However, the color tuning mechanism that leads to firefly multicolor bioluminescence is still unknown, nor is consensual which is the yellow-green and red emitters. Computational studies have been essential in the study of oxyluciferin (OxyLH₂) chemi- and bioluminescence and are responsible for most of our knowledge of this natural phenomenon. The objective of this manuscript is the analysis of the benefits and the conclusions derived from the theoretical studies of the light emitter, OxyLH₂, and its applications on bioluminescence research.

INTRODUCTION: THE FIREFLY MULTICOLOR BIOLUMINESCENCE ENIGMA

The emission of light resulting from an enzyme catalyzed biochemical reaction is known as bioluminescence. This natural phenomenon is found in many types of organisms, including fungi, insects, bacteria, worms, dinoflagellates, and fish. Despite some structural differences, the majority of these bioluminescence reactions are catalyzed by an enzyme named luciferase, which reacts with different substrates called luciferin.^{1,2} In recent years, this bioluminescence system has gained numerous bioanalytical, biomedical, and pharmaceutical applications, among others. More specifically, it is involved in the analytical determination of adenosine 5'-triphosphate (ATP) in microbial detection, immunoassays, bioimaging, biosensing and is used as a gene reporter.^{1–7}

The most studied bioluminescent reaction known is that of the North American firefly *Photinus pyralis*.^{1,3} *Photinus pyralis* luciferase (EC 1.13.12.7, Luc) catalyzes a two-step reaction: The first is the condensation reaction between the luciferin substrate (LH₂), a derivative of benzothiazolyl-thiazole, and ATP in the presence of Mg²⁺. The second step consists of the oxidation of the first step product, an adenylyl intermediate (LH₂-AMP) and the release of AMP, CO₂, and OxyLH₂. The light emitter is formed in an excited singlet state S₁, which decays to the ground state with the emission of visible light (Scheme 1). This system is known for its efficiency when compared with chemiluminescence. For many years the efficiency of this reaction was thought to be 88%,^{8,9} but a recent work performed by Ando et al. estimated it at 41%.¹⁰ Albeit the sharp decrease, this new value still strongly supports the study and the development of practical applications for this bioluminescence system. Other firefly luciferase system commonly studied include *Luciola cruciata*, *Luciola lateralis*, *Luciola mingrelica*, *Phrixotrix hirtus*, *Lampyrus noctiluca*, and *Lampyrus turkestanicus*.^{1,11}

Currently, one of the most intriguing and studied aspects of firefly light emission is the origin of the multicolor bioluminescence. Albeit the similarity of the substrate–product structures between all bioluminescent insects, their emission energies range

from 2.14 to 2.34 to 2.00 eV.^{12,13} The red shift can be induced by high temperatures, addition of divalent metal cations, and denaturation by addition of substances like urea.^{14–16} The shifting from yellow-green (basic pH) to red emission is also achieved with a decrease in pH.⁹ The understanding of this peculiar aspect could be of enormous importance in firefly research. Red-emitting Luc could be used in *in vivo* medical imaging, as red light is absorbed very poorly by mammalian tissues in comparison with the natural emitted light. Also, the control of the multicolor bioluminescence could be the basis for using Luc as a single dual reporter gene, a bioindicator of cellular stress, and a probe for intracellular changes of pH.¹⁷

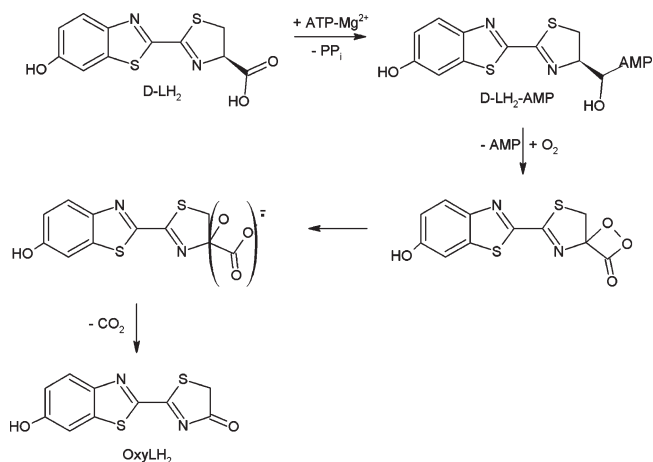
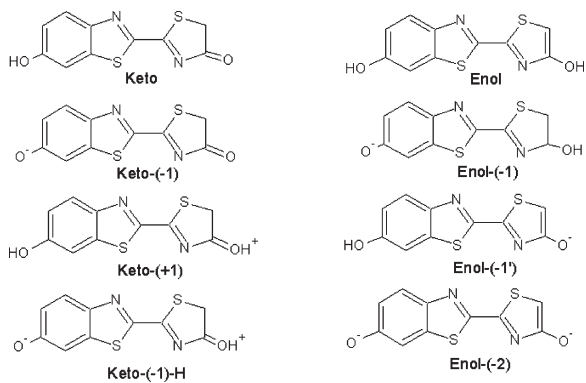
Several hypotheses have been proposed over the years to try to explain this pH-dependent phenomenon:

- The first hypothesis was advanced by White et al. who suggested that the color variation was the result of the keto–enol tautomerism of the OxyLH₂ species (Scheme 2).¹⁸ The keto form was supposed to emit red light, while the enol form was the yellow-green emitter. However, experiments with a keto-constrained OxyLH₂ analogue indicated that the light-emitting reaction only required a keto emitter.¹⁹
- McCapra et al.²⁰ proposed a mechanism based on theoretical calculations with the semiempirical functional AM1.²¹ These calculations indicated that the color variation depended on the rotation around the C–C bond of the –N=C–C=N– moiety.
- The dependence of the bioluminescence color on the polarization of the internal microenvironment of Luc was advanced by several groups as a determining factor in the light-emitting mechanism. The higher the polarizability, the larger the red shift of bioluminescence.^{4,22,23}
- Following its 2002 work,¹⁷ Branchini et al. proposed that Luc modulates the color of the light emitted by controlling the resonance structure of the anionic keto form of

Received: January 3, 2011

Published: March 07, 2011

Scheme 1. Luc-Catalyzed Bioluminescence Reaction

Scheme 2. Tautomeric and Dissociation States of OxyLH₂

OxyLH₂.²⁴ A structure showing a $-\text{N}=\text{C}=\text{C}=\text{N}-$ moiety would relax by emitting green light, while another one showing a $=\text{N}-\text{C}=\text{C}-\text{N}=-$ moiety would be the red emitter.

- (e) Nakatsu et al. proposed another mechanism based on the recently solved three-dimensional structure of *Luciola cruciata* luciferase (LcLuc).²⁵ This structure and its mutant showed that this enzyme can assume two different conformations: an open conformation with a polar and loose active site and a closed conformation with a hydrophobic and rigid active site. Based on these findings, this group hypothesized that the geometrical rearrangement of the active site could lead to different relaxation of the light emitter. So, a nonrelaxed OxyLH₂ would emit yellow-green light, and a molecule that is permitted to relax would emit red light.
- (f) More recently Hirano et al. proposed a mechanism derived from hypothesis (c).²⁶ They stated that the emission of light is modulated by the polarity of the active site internal environment and by the degree of covalent character of a bond between the anionic keto form and a protonated basic moiety present in the active site.

Thus, these proposed mechanisms are all based on changes in the properties of OxyLH₂, and/or changes in the internal microenvironment of the active site. However, there is still no

consensus on which hypothesis best describes the mechanism behind the multicolor bioluminescence. One factor that impairs the experimental study of this question is the experimental instability of OxyLH₂ in a basic model solution. In fact, only rather recently, the crystal structure of this molecule was achieved for the first time.²⁷ Therefore, the majority of information was obtained from more stable analogues, as 5-methyl (MOxyLH₂), 5,5-dimethyl (DMOxyLH₂), and 6'-dehydroxy (DHOxyLH₂).^{28–32} However, studies with these molecules only provide indirect information as they possess substitutions which impose significant steric hindrance for binding of the emitter to Luc, and DMOxyLH₂ is keto-constrained and cannot account for possible effects related to the keto–enol tautomerism. Moreover, these studies are performed outside the protein active center.

Thus, a computational approach to the firefly bioluminescence research is fundamental. Only computational studies appear to have the required level of accuracy and reliability for understanding this natural phenomenon. One of the main advantages of the use of theoretical calculations is the possibility of direct study of the properties of the various OxyLH₂ species in various solvents, without the need for more stable analogues. Also no experimental technique, unlike the available theoretical methods, allows us to perform direct and detailed studies, down to the atomistic level, of the interactions between the emitter molecule with molecules present in the active site of Luc and the effect of these interactions on the emission energies.

THEORETICAL STUDIES OF OXYLH₂ STRUCTURAL AND ELECTRONIC PROPERTIES

The light emitter OxyLH₂ is formed in the singlet excited state in the Luc-catalyzed reaction by attachment of O₂ to the thiazolone moiety of LH₂–AMP. This reaction releases AMP and forms a dioxetanone intermediate which decarboxylates into the emitter molecules.¹ Besides bioluminescence, OxyLH₂ also exhibits chemiluminescence. It was established that this molecule is in the keto form in dimethyl sulfoxide (DMSO) solution containing a small amount of potassium *tert*-butoxide, and it is transformed into the enol form in DMSO with a large amount of potassium *tert*-butoxide.^{18,33,34} The enol form was thought to emit green light and the keto form red light.¹⁸ It was this similarity between chemi- and bioluminescence which introduced the idea that the color tuning mechanism was the same in the two processes.^{18,33} That is why that the first computational studies on bioluminescence focus on OxyLH₂ intrinsic properties to explain the multicolor bioluminescence. The first relevant study was performed by McCapra et al.²⁰ It was made in vacuo and by means of the semiempirical functional AM1 and proposed the twisted intramolecular charge-transfer (TICT) state mechanism. According to their calculations, the yellow-green emitter has a planar structure and is a saddle point on the potential energy surface. The twisting between the $-\text{N}=\text{C}=\text{C}=\text{N}-$ stabilizes the molecule and leads to red emission. However, the method used has a tendency not to describe conjugated single bonds and underestimated the energy barrier between these two states.^{35,36} This was later confirmed by another in vacuo study which used the more reliable *ab initio* excited-state method,³⁷ configuration interaction with single excitations (CIS).³⁸ These calculations predicted that the planar structure of both the enol and keto forms is a minimum on the S₁ potential energy surface and that the twisted structure is a saddle point.

Table 1. Absorption Energies (in eV) of the Various OxyLH₂, Calculated at Different Levels of Theory^a

OxyLH ₂	TD-B3LYP/6-31+G(d)	TD-B3LYP/6-31G(d)	TD-B3LYP/6-31+G(d,p)	CIS/6-31G(d)
keto-(3.35) ^{c,d}	3.32 ^{b,37} 3.02 ^{c,63} 3.32 ^{b,64}	3.55 ^{b,51} 3.42 ^{c,51}	3.32 ^{b,43}	5.15 ^{b,51} 5.07 ^{c,51}
enol	3.40 ^{b,37} 3.20 ^{c,63}	3.71 ^{b,51} 3.66 ^{c,51}		4.95 ^{b,51} 4.89 ^{c,51}
keto-(−1) (2.99) ^{c,e}	2.54 ^{b,37} 2.45 ^{c,63} 2.54 ^{b,64} 2.55 ^{c,64}	2.72 ^{b,51} 2.75 ^{c,51}	2.54 ^{b,43}	3.49 ^{b,51} 3.55 ^{b,51}
enol(−1)	2.48 ^{b,37} 2.54 ^{c,63}	2.65 ^{b,51} 2.77 ^{c,51}		3.58 ^{b,51} 3.81 ^{c,51}
enol(−1')	2.06 ^{b,37}	2.01 ^{b,51} 2.41 ^{c,51}		3.74 ^{b,51} 4.16 ^{c,51}
enol(−2)	2.36 ^{b,37} 2.49 ^{c,63}	2.58 ^{b,51} 2.67 ^{c,51}		4.22 ^{b,51} 4.22 ^{b,51}
keto(+1)	2.40 ^{c,63} 2.27 ^{b,64} 2.47 ^{c,64}			
keto(−1)-H	2.39 ^{b,64} 2.41 ^{c,64}			

^a Experimental values are in parentheses. ⁵⁰ ^b Gas phase. ^c Water. ^d pH 1.0. ^e pH 10.0.

Table 2. Emission Energies (in eV) of the Various OxyLH₂, Calculated at Different Levels of Theory^a

OxyLH ₂	TD-B3LYP/6-31+G(d)	TD-B3LYP/6-31G(d)	SAC-CI	CIS/6-31G(d)
keto	2.71 ^{c,63} 2.99 ^{b,64}	3.17 ^{b,51} 3.12 ^{c,51}	2.95 ^{d,57}	4.13 ^{b,51} 4.03 ^{c,51}
enol (2.77) ^d	2.73 ^{c,63}	3.07 ^{b,51} 3.03 ^{c,51}	2.83 ^{d,57}	3.77 ^{b,51} 3.72 ^{c,51}
keto(−1) (2.01) ^d	2.20 ^{c,63} 2.23 ^{b,64} 2.30 ^{c,64}	2.61 ^{b,51} 2.63 ^{c,51}	2.08 ^{d,57}	3.34 ^{b,51} 3.40 ^{c,51}
enol(−1) (2.22–2.24, ^c 2.16) ^d	2.44 ^{c,63}	2.59 ^{b,51} 2.65 ^{c,51}	2.25 ^{d,57}	3.22 ^{b,51} 3.33 ^{c,51}
enol(−1')		1.97 ^{b,51} 2.19 ^{c,51}	2.14 ^{d,57}	2.89 ^{b,51} 3.22 ^{c,51}
enol(−2) (2.30, ^c 2.10) ^d	2.17 ^{c,63}	2.19 ^{b,51} 2.27 ^{b,51}	2.07 ^{d,57}	3.14 ^{b,51} 3.17 ^{c,51}
keto(+1)	2.05 ^{c,63} 1.70 ^{b,64} 2.10 ^{b,64}			
keto(−1)-H	1.79 ^{b,64} 1.99 ^{c,64}			

^a Experimental values are in parentheses. ²⁷ ^b Gas phase. ^c Water. ^d DMSO.

The paper of Orlova et al. was the first work to give a theoretical insight on the stability of the different OxyLH₂ species and their vertical excitation energies.³⁷ Theoretical calculations with the B3LYP functional³⁹ and 6-31+G(d) split-valence basis set augmented with *d*-type polarization and diffuse functions⁴⁰ predicted that the anionic keto-trans conformer is more stable than keto-cis by 5.3 kcal/mol. This was an important breakthrough as the earlier modeling studies of the active site assumed a cis-conformation for LH₂ and OxyLH₂.^{41,42} This study was complemented by calculations (B3LYP/6-31+G(d,p)) of Liu et al., which predicted that all of the trans-conformers are more stable than the cis species by a range of 4.9–6.2 kcal/mol.⁴³ Orlova et al. also predicted that enol and enol(−1) are less stable by 8.5 and 19.9 kcal/mol than their corresponding keto forms.³⁷ This was another pivotal discovery because combined with Branchini et al. conclusions, directed the research of firefly bioluminescence for the study of the keto forms of OxyLH₂.¹⁹ Another important contribution from this work was the first study of the in vacuo vertical excitation energies of the OxyLH₂, which are summarized in Table 1. For this, excited-state predictions were made by means of time-dependent (TD) hybrid time density functional theory (DFT) method (TD-B3LYP)^{44,45} and, for additional calibration, the Zerner's intermediate neglect of differential overlap (ZINDO) semiempirical method.^{46,47} Both methods were already proven reliable in some excited-state calculations.^{48,49} The results obtained with TD-B3LYP gave excellent agreement with experiment at pH = 1.⁵⁰ This opened the way for the use of TD-DFT in the study of OxyLH₂ electronic properties.

The conclusions regarding bioluminescence that could be derived from this study are however limited, as aqueous solvents always play an important role in vivo and can affect significantly excited-states energies. Ren and Goddard were the first to include

solvent effects in their vertical excitation energies calculations (Table 1).⁵¹ Moreover, they were the first to perform the systematic excited-state optimizations of OxyLH₂ and the calculation of the corresponding emission energies (Table 2). The excited-state calculations were made with the TD-B3LYP and CIS methods with the 6-31G(d) basis set. It should be noted that the basis set used lacked diffuse functions, which are necessary for the accurate prediction of excited-state energies.³⁷ The predictions of solvent effects were made with the self-consistent isodensity polarized continuum model (SCI PCM) with parameters set for water.⁵² The obtained results showed the importance of solvent effects on the prediction of the excitation and emission energies of the anionic species, as they underwent large shifts to the blue when in comparison with in vacuo results. The neutral species were affected to a lesser extent and suffer only small shifts to the red. The CIS method correctly predicted the changes in the excited-state geometries but needed a scaling of the wavelengths to be in reasonable agreement with the TD-B3LYP results and with experiment. This indicated that calculations with the CIS method should be limited to geometries optimization and that the prediction of OxyLH₂ electronic properties should be made with more accurate computational methods.

Goddard's group continued the theoretical study of the S₁ state of OxyLH₂ by performing multireference calculations.⁵³ They used the complete active space self-consistent-field (CASSCF) method⁵⁴ and the multiconfigurational complete active space second-order perturbation theory (CASPT2) in the in vacuo study of the structural and electronic properties of keto(−1) and enol(−1).^{55,56} The CASSCF method can be used to predict accurate ground- and excited-states structures, and CASPT2 is used to include dynamic electron correlation corrections, which can be useful for obtaining reliable excitation and emission energies. These methods are considered more accurate and

reliable than the CIS and TD-DFT methods. The group confirmed Orlova et al. results by showing that planar keto(-1) and enol(-1) are minima on both the S_0 and the S_1 potential energy surfaces and that the twisted forms are transition states.³⁷ They also predicted emission energies in the range of 2.35–2.53 eV for keto(-1) and 2.42–2.74 eV for enol(-1). The strong oscillator strengths predicted are consistent with a strong S_1 – S_0 vertical emission.

Nakatani et al.⁵⁷ used the symmetry adapted cluster⁵⁸/symmetry adapted cluster–configuration interaction (SAC/SAC–CI)⁵⁹ method for the study of OxyLH₂.^{60,61} This method can be used for balanced description of the electron correlation effects in both ground and excited states. One of the objectives delineated by this group was the study of the chemiluminescence of OxyLH₂ in DMSO, as described in refs 18,33, and 34. To simulate the DMSO environment, a polarized continuum model (PCM) was used, with the dielectric constant of 46.7.⁶² Their calculation of the emission energies (Table 2) characterized the neutral species as blue emitter, excluding them from the candidates for the emitters. The emission energy of keto(-1) agreed well with red emission, while the anionic enol forms all emit in the yellow-green region. However, enol(-2) was predicted to be more stable by 6.6–8.3 kcal/mol than the other species, characterizing it as the yellow-green emitter. It should be noted that the SAC–CI method was used in single point calculations, while the optimizations were performed with the CIS method. This emphasized the reliability of this method in geometry optimization calculations.

The recent use of more accurate ab initio methods led to the abandonment of TD-DFT methods in OxyLH₂ research.^{43,53,57} These methods were being criticized for its predictions on OxyLH₂ due to possible charge-transfer (CT) states, for which TD-DFT shows large errors.^{53,61} However, Li et al. showed that CT was large for the twisted forms of OxyLH₂ but small for the planar ones, devaluing this flaw of TD-DFT methods in firefly bioluminescence research.⁶³ This group also tried to study the pH-dependent fluorescence spectra of OxyLH₂ in aqueous solution, by means of TD-B3LYP/6-31+G(d). Solvent effects were treated with the PCM model, with parameters set for water. Their results (Table 2) attributed the blue fluorescence peak (450 nm, 8 > pH > 3) to the neutral keto and enolic species and the yellow-green (560 nm, pH > 9) to enol(-1). To explain the red peak (620 nm, pH < 3), a new species keto-(+1) was considered.

■ MICROSOLVATION STUDIES OF MULTICOLOR BIOLUMINESCENCE

The studies presented so far provided valuable information regarding OxyLH₂ structural and electronic properties and gave us some insights about its chemiluminescence. However, these studies did not take into account the microenvironment of Luc active site, and some of these studies do not take into account solvent effects of any kind. Due to the complexity of a protein active site and the various interactions that its molecules can make with the protein substrate, it is not likely that the effect of Luc on the emission can be disregarded or minimized. Therefore, these studies are limited to chemiluminescence and some indirect information of the firefly light-emission phenomenon.

Some authors have begun to address this problem by creating more complex simulations, which are focused on some aspects of Luc microenvironment. However, as the size of Luc–OxyLH₂ is

incompatible with the most accurate quantum mechanics methods, more simplified models are still used and can provide important information. Liu et al. studied the effect of the polarization of the microenvironment on the emission energies by connecting a H₂O or a CH₂Cl₂ molecule to keto to simulate solvents of different polarity, at the B3LYP/6-31+G(d,p) level.⁴³ The in vacuo TD-B3LYP/6-31+G(d,p) calculated excitation energies (Table 1) decreased on the order of keto, keto-CH₂Cl₂, keto-H₂O, and keto(-1), which in their model above corresponded to an increase in the polarization of the microenvironment. This is consistent with hypothesis (c) described above. However, it should be noted that an accurate study of the effect of the polarization of the microenvironment should be made including the crucial solvent effects, which are disregarded in this study. According to Ren and Goddard results, the inclusion of solvent effects could provoke a large blue shift of keto(-1), when comparing with the in vacuo results, and a small red shift of keto and keto-X complexes.⁵¹ These possible shifts could suffice for altering the excitation energies ordering presented by Liu et al.⁴³ Moreover in our opinion, the addition of these molecules to keto simulates the interactions between this species with molecules present in the Luc active site rather than simulating different solvents. This indicates that the addition of CH₂Cl₂ or H₂O to keto(-1) could provoke different effects on its emission energies than caused in the case of keto and so change this descending order. This group also demonstrated that keto(-1), which had excitation energies (2.54 eV) closer to experiment, has a flat potential energy surface which allows an easy shifting of the minimum between different resonance structures by means of CASSCF geometry optimizations and multistate CASPT2 single point calculations. They also showed that the relaxation of keto(-1) on S_1 can change significantly the emission energies. In conclusion, they stated that the hypotheses (c–e) are all plausible.

Min et al. tried to study the role of the resonance structure of keto(-1) on the multicolor bioluminescence by performing a TD-DFT investigation on the origin of the red chemiluminescence.⁶⁴ They connected a sodium or an ammonium cation to the benzothiazole or the thiazolone oxygen of keto(-1) and studied their absorption and emission energies with the B3LYP, B3PW91, and PBE1KCIS functionals and the 6-31+G(d) basis set.^{39,65,66} The PCM model and the conductor-like screening model (COSMOS) were used to simulate an aqueous environment.⁶⁷ They demonstrated that the interactions of keto(-1) with the two cations caused similar blue shifts, while connected to the benzothiazole oxygen, and similar red shifts when interacting with the thiazolone moiety. These opposite effects caused by different interactions between the same molecules emphasizes the importance of the various interactions that can be formed in the complex active site microenvironment between the light emitter and the molecules present in Luc active sites. In this work it was considered a novel emitter, keto(-1)-H. This new species could be of some importance in the bioluminescence phenomenon due to its emission wavelength (624.1–650.7 nm), which is close to the experimental value of 620 nm at acid pH.¹

In order to assess the role of the rigidity of Luc active site on light emission, as described by Nakatsu et al.,²⁵ Li et al. focused on the study of the excited-state geometry of keto(-1).⁶⁸ According to this group, the emitter, following changes in the electronic structure from a initial excited state on the potential energy surface, usually relaxes to a energy minimum and emits a photon while returning to the ground state. The different

conformations of Luc could affect the energy values of these points in the potential energy surface, modulating the color of bioluminescence. Their B3LYP/6-31+G(d) and TD-B3LYP/6-31+G(d) calculations demonstrated that changes in six bond lengths of keto(-1) excited-state geometry can determine the emission spectra. The more the bond lengths changes are impeded, the more the emission energies increase. It should be noted, however, that this study was performed without imposing any constraints on these geometry changes. Due to the high complexity of interactions between the molecules present in protein active site and to the steric constraints imposed by the active site to the substrate caused by the proximity to active site molecules, it is expected that Luc has a great influence on bond lengths changes of OxyLH₂ and is unlikely that the emitter has the necessary flexibility to suffer so many significant changes in its geometry.

Cai et al. calculated the excitation energies of keto(-1) in response to different electrostatic fields computed using TD-DFT functionals.⁸⁴ They found the existence of a correlation between the wavelength shift with the projection of the electrostatic field on the molecular plane and the intensity of fluorescence can be affected by field modulation. However, neither the study of the effect of polarity in these electrostatic fields nor the integration of this simple model in the in vitro pH-dependent color variation were performed. Also, in this paper the use of the local density approximation (LDA) was validated, a much less computational demanding functional than B3LYP which is used in numerous optical calculations.^{85,86}

■ INTEGRATION OF LUC ENVIRONMENT IN BIOLUMINESCENCE RESEARCH

Apart from these simplified models, the latest tendency in the theoretical bioluminescence research is the incorporation of extended portions of the active site in the calculations. Despite the valuable information that can be derived from more simplistic simulations, these studies ignore the steric and electrostatic contributions from Luc active site. Thus, some groups have recently begun to study the active site contributions to the light-emitting reaction. The first study in this field was made by Nakatani et al.⁵⁷ They performed quantum (QM) and molecular mechanical (MM) calculations, based on the X-ray structure of Luc and some working models derived from experimental studies, to determine the structure of the excited-state Luc-OxyLH₂ complex.^{41,42,69} This group stated that the Luc environment shifted the emission energy of keto(-1) to the green region (2.08–2.33 eV) and that Arg218, His245 (Luc numbering), and the phosphate group of AMP gave dominant contributions to this effect. It was also stated that the anionic enol species have emission energies close to experiment, but the keto-enol tautomerism is energetically unfavorable in Luc environment. The resonance-based mechanism was dismissed by these authors, as they did not find significant changes in the resonance structure in the excited state. However, some caution is needed in the analysis of this conclusion. The authors reached to this dismissal by comparing the bond lengths of OxyLH₂ in the gas phase and in their model. However, by Nakatsu et al. findings, Luc can adopt either a tight and hydrophobic active site or a loose and polar one.²⁵ Thus, these differences in Luc internal micro-environment can suffice for changes in the resonance structure of OxyLH₂. This group continued their studies in bioluminescence by performing in silico mutagenesis SAC-CI experiments.⁷⁰ By

analyzing the contributions of several amino acid residues to the emission color tuning, they demonstrated the blue-shift effect of Arg223, Glu344, and Asp422. The replacement of these amino acid residues by an alanine caused a red-shift, thus predicting potential targets for future mutagenesis studies.

Tagami et al. performed a similar study on LcLuc bioluminescence.⁷¹ They employed the multilayer fragment molecular orbital method in combination with CIS(D) calculations^{72,73} in the study of the emission energies resulting from the interaction between OxyLH₂ and the wild-type and mutant crystal structures determined by Nakatsu et al.²⁵ The experimental results were not well reproduced but were improved by the use of the whole structure of the enzymes in the calculations. These results further emphasize the importance of Luc contribution for the color tuning mechanism.

Another QM/MM investigation was conducted by Navizet et al., which was based on the open and closed conformation of LcLuc.⁷⁴ In this work, the authors create several models of the open and closed conformations of LcLuc structure. The main differences between the models were the number of the water molecules present in the active site and the performance of extensive molecular dynamics simulation on one of the models. This work has special importance on the multicolor bioluminescence research, as it is the only one to take into account the different conformations of Luc active site in more complex calculations. Their CASPT2/CASSCF calculations on keto(-1) demonstrated that the polarization of the microenvironment of the benzothiazole moiety have a crucial effect on the light emission. Moreover, the results obtained were in disagreement with the experimental conclusion that the rigidity of the active site controls the bioluminescence color and reduces the role of Luc different conformations on light emission to the modulation of the polarity of the microenvironment. It should be noted that these pivotal conclusions were achieved with a noteworthy reproduction of the spectral parameters of light production.

Three of the latest computational studies on firefly bioluminescence also focus on the contribution of the active site molecules to the multicolor variation. Milne et al. use the fragment molecular orbital method to study the effect of some amino acid residues, water molecules, and AMP on the excitation energies of some OxyLH₂ species.⁷⁵ This study assumes some importance as it is the only one to perform a systematic analyses of the effect of an extended portion of the active site on more OxyLH₂ species than keto(-1). Moreover, it introduces a sense of pH variation on the simulation by considering different protonation states for AMP. Based on their calculations, the group proposed that keto(-1) is the yellow-green and red emitter. However, this conclusion was reached by a somewhat confusing comparison between their calculated excitation energies and the experimental emission energies. Furthermore, the calculations were based on the assumption that AMP has a pK_a value of 6.23, which due to the variable internal environments of enzymatic active site may not be true in the case of Luc.⁷⁶

Min et al. continued their TD-DFT-based studies on firefly bioluminescence, by constructing a complex between keto(-1), AMP, and some other important active site molecules in order to simulate Luc contribution to the yellow-green bioluminescence.⁷⁷ This study gains importance relative to others here described, as it is the only to include implicit solvent effects by means of the COSMO model with parameters set for water, in their TD-(B3LYP, B3PW91, PBE1KCIS)/6-31+G(d) calculations. The results obtained indicate that keto(-1) is the yellow-green

emitter, which is consistent with Nakatani et al. and Milne et al.^{57,75} However, some doubt is shed on these results by the use of water to simulate the environment of yellow-green emission. It is unlikely that the polar environment of an aqueous solution could be used in a reliable simulation of a hydrophobic active site, as predicted by Nakatsu et al.²⁵ Moreover, in the literature is described that for active site simulations a dielectric constant of 4 gives good agreement with experiment, which has large differences relative to the dielectric constant of water (~ 78).^{78–81} Therefore, it is reasonable to speculate if the emission energies of keto(-1)-Luc would suffer an undesirable shift, when used with a proper dielectric constant, or if the polarity of the different conformations of Luc do not have any significant impact on bioluminescence.

Mao studied the sequence-induced color variation of Luc by performing dynamics simulations.⁸² An elastic network model was used with a coarse-grained representation of protein and a harmonic potential describing the interactions of its components.⁸³ This work permitted the identification of several hotspot residues that are coupled to the active site and identified the B subdomain as being mainly responsible for the multicolor variation. However, this study was based on certain assumptions that clash with previous knowledge regarding firefly bioluminescence. First, the Luc structures considered for this work were unbound Luc as the open conformation and LcLuc bound to AMP and OxyLH₂ as the closed conformation, while most studies considered LcLuc bound to DLSA as the closed conformation and LcLuc bound to AMP and OxyLH₂ as the open conformations.^{25,69} Moreover, the hotspots were identified by measuring changes in the mean-square fluctuations of the amino acid residues, a measure that in the opinion of the author may indicate the impact of the residues in the function of the protein. However, as the closed conformation here employed is very similar to the structure of LcLuc bound to ATP in the beginning of the bioluminescence reaction and no excitation and/or emission energies calculations were performed, it is not evident that the function modulated by these residues is the emission of light or the adenylation step that initiates this reaction.²⁵

■ FUTURE PERSPECTIVES ON COMPUTATIONAL STUDIES OF BIOLUMINESCENCE

Computational calculations have been a powerful tool in the research of the OxyLH₂ molecule. The studies performed to date have given us valuable insight on the species structural and electronic properties and its role on chemiluminescence in solvents like water and DMSO, without the need for more stable analogues.

The use of a computational approach in the study of firefly bioluminescence has also been recurrent but with less conclusive results. The use of computational techniques emphasized the importance of Luc in the light-emitting phenomenon and has provided qualitative analyses for the contribution of some key amino acid residues, which could be useful in mutagenesis studies. However, there is still some controversy and lack of substantiated information on other aspects of the bioluminescence phenomenon.

For example, the latest theoretical studies are considering keto(-1) as the yellow-green emitter. However most authors, based on Branchini et al. results, already assume that keto(-1) is the most probable emitter and do not study the enolic species in their simulations.¹⁹ It should be noted that experiment made with the keto analogue only demonstrated that the keto species

could produce yellow-green bioluminescence, and no experimental evidence excluded the enolic species from bioluminescence. However, some clarification to this topic may be provided by two very recent papers. Navizet et al. performed an analysis on six OxyLH₂ chemical forms (keto, enol, and the respective anions) using a multireference method.⁸⁷ Their MS-CASPT2 calculations in vacuo and in DMSO excluded keto, enol, and enol(-1') as possible light emitters, while MS-CASPT2/MM calculations on the remaining species indicated that keto(-1) is the direct excited-state product of firefly dioxetanone and the sole light emitter. Also, our own group studied the effect of pH on the chemical equilibrium of OxyLH₂ (keto(-1)-H, keto, enol, and respective anions) in the open and closed conformations of the Luc active site.⁸⁸ To this end, we simulated solvent effects by using the CPCM model with two different dielectric constants (4 and 78). Our TD-PBE0⁸⁹ calculations indicated that keto(-1) is the sole species present in both conformations at the pH range of interest.

Furthermore, the majority of the studies do not take into account changes in the polarity of the active site, as described by Nakatsu et al.,²⁵ when they are studying the red-shift of the color of emitted light. Moreover, most of the studies described here do not take into account solvent effects of any kind, focusing instead on in vacuo calculations. This could lead to erroneous results as Ren and Goddard and Min et al. described the significant differences in the emission energies that can arise from when we compare results obtained in the presence or absence of solvent.^{51,64} This leads to the questions: if solvent effects are considered, then will keto(-1) emission energies still agree well with experiment? If not, then which is the real yellow-green emitter? If yes, then does the solvent polarity affect the bioluminescence color? To what extent does the polarity of the micro-environment affect firefly bioluminescence?

There is also some lack of knowledge regarding the active site of Luc that can prevent obtaining reliable results from the more complex models. For example, there has been some discrepancy in the protonation state of a histidine residue located near the thiazolone moiety of the emitter. The variation of the protonation state of AMP was considered pivotal in the red shift, but no study was made in order to validate this variation. Moreover no study considered, at least explicitly, a proton acceptor for the C₄ proton of LH₂.¹ This could be of great importance, as it could generate an unexpected protonation state and cause a rearrangement of the internal interaction of the active site.

Computational studies have been undeniably fundamental in building our knowledge of the firefly multicolor bioluminescence. Thus, it could be of pivotal importance in future lines of research, more specifically, on the study of contributions of all OxyLH₂ species to the color tuning mechanism, the study on Luc microenvironment contribution with solvent effects, the correct characterization of Luc active site, and the study of OxyLH₂ formation.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: jcsilva@fc.up.pt.

■ ACKNOWLEDGMENT

Financial support from Fundação para a Ciência e a Tecnologia (FCT, Lisbon) (Programa Operacional Temático Factores de Competitividade (COMPETE) e participado pelo Fundo

Comunitário Europeu FEDER) (Project PTDC/QUI/71366/2006) is acknowledged.

REFERENCES

- (1) Marques, S. M.; Esteves da Silva, J. C. G. Firefly bioluminescence. A mechanistic approach of luciferase catalyzed reactions. *IUBMB Life* **2009**, *61*, 6–17.
- (2) Leitão, J. M. M.; Esteves da Silva, J. C. G. Firefly luciferase inhibition. *J. Photochem. Photobiol. B* **2010**, *101*, 1–8.
- (3) Inouye, S. Firefly luciferase: an adenylate-forming enzyme for multicatalytic functions. *Cell. Mol. Life Sci.* **2010**, *67*, 387–404.
- (4) Ugarova, N. N. Interactions of firefly luciferase with substrates and their analogues: a study using fluorescence spectroscopy methods. *Photochem. Photobiol. Sci.* **2008**, *7*, 218–227.
- (5) Meroni, G.; Rajabi, M.; Santaniello, E. D-luciferin, derivatives and analogues: synthesis and *in vitro/in vivo* luciferase-catalysed bioluminescence activity. *ARKIVOC* **2009**, 265–288.
- (6) Luker, K. E.; Luker, G. D. Applications of bioluminescence imaging to antiviral research and therapy: multiple luciferase enzymes and quantitation. *Antiviral Res.* **2008**, *78*, 179–187.
- (7) Fan, F.; Wood, K. V. Bioluminescence assays for high-throughput screening. *Assay Drug Dev. Technol.* **2007**, *5*, 127–136.
- (8) Seliger, H. H.; McElroy, W. D. Spectral emission and quantum yield of firefly bioluminescence. *Arch. Biochem. Biophys.* **1960**, *88*, 136–141.
- (9) Seliger, H. H.; McElroy, W. D. Colors of firefly bioluminescence – enzyme configuration + species specificity. *Proc. Natl. Acad. Sci. U.S.A.* **1964**, *52*, 75–81.
- (10) Ando, Y.; Niwa, K.; Yamada, N.; Enomoto, T.; Irie, T.; Kubota, H.; Ohmiya, Y.; Akiyama, H. Firefly bioluminescence quantum yield and colour change by pH-sensitive green-emission. *Nat. Photonics* **2008**, *2*, 44–47.
- (11) Fraga, H. Firefly luminescence. A historical perspective and recent developments. *Photochem. Photobiol. Sci.* **2008**, *7*, 218–227.
- (12) Wood, K. V. The chemical mechanism and evolutionary development of beetle bioluminescence. *Photochem. Photobiol.* **1995**, *62*, 662–673.
- (13) Hastings, J. W. Chemistries and colors of bioluminescent reactions: A review. *Gene* **1996**, *173*, 5–11.
- (14) Seliger, H. H.; Buck, J. B.; Fastie, W. G.; McElroy, W. D. Spectral distribution of firefly light. *J. Gen. Physiol.* **1964**, *48*, 95–104.
- (15) Lee, R. T.; Denburg, J. L.; McElroy, W. D. Substrate-binding properties of firefly luciferase 0.2. ATP-binding site. *Arch. Biochem. Biophys.* **1970**, *141*, 38–52.
- (16) Lundovskikh, I.; Dementieva, E.; Ugarova, N. Recombinant firefly luciferase in *Escherichia coli* – Properties and immobilization. *Appl. Biochem. Biotechnol.* **2000**, *88*, 127–136.
- (17) Viviani, V. R.; Arnoldi, F. G. C.; Neto, A. J. S.; Oehlmeier, T. L.; Bechara, E. J. H.; Ohmiya, Y. The structural origin and biological function of pH-sensitivity in firefly luciferases. *Photochem. Photobiol. Sci.* **2008**, *7*, 159–169.
- (18) White, E. H.; Rapaport, E.; Seliger, H. H.; Hopkins, T. A. The chemi- and bioluminescence of firefly luciferin: An efficient chemical production of electronically excited states. *Bioorg. Chem.* **1971**, *1*, 92–122.
- (19) Branchini, B. R.; Murtiashaw, M. H.; Magyar, R. A.; Portier, N. C.; Ruggiero, M. C.; Stroh, J. G. Yellow-green and Red Firefly Bioluminescence from 5, 5-Dimethyloxyluciferin. *J. Am. Chem. Soc.* **2002**, *124*, 2112–2113.
- (20) McCapra, F.; Gilfoyle, D. J.; Young, D. W.; Church, N. J.; Spencer, P. In *Bioluminescence and Chemiluminescence: Fundamental and Applied Aspects*; Campbell, A. K., Kricka, L. J., Stanley, P. E., Ed.; Wiley: New York, 1994; pp 387–391.
- (21) Dewar, M. J. S.; Zois, E. G.; Healy, E. F. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.
- (22) Morton, R. A.; Hopkins, T. A.; Seliger, H. H. Spectroscopic properties of firefly luciferin and related compounds, and approach to product emission. *Biochemistry* **1969**, *8*, 1598–1607.
- (23) DeLuca, M. Hydrophobic nature of active site of firefly luciferase. *Biochemistry* **1969**, *8*, 160–166.
- (24) Branchini, B. R.; Southworth, T. L.; Murtiashaw, M. H.; Magyar, R. A.; Gonzalez, S. A.; Ruggiero, M. C.; Stroh, J. G. An Alternative Mechanism of Bioluminescence Color Determination in Firefly Luciferase. *Biochemistry* **2004**, *43*, 7255–7262.
- (25) Nakatsu, T.; Ichiyama, Y.; Hiratake, J.; Saldanha, A.; Kobashi, N.; Sakata, K.; Kato, H. Structural basis for the spectral difference in luciferase bioluminescence. *Nature* **2006**, *440*, 372–376.
- (26) Hirano, T.; Hasumi, Y.; Ohtsuka, K.; Maki, S.; Niwa, H.; Yamaji, M.; Hashizume, D. Spectroscopic Studies of the light-Color Modulation Mechanism of Firefly (beetle) Bioluminescence. *J. Am. Chem. Soc.* **2009**, *131*, 2385–2396.
- (27) Naumov, P.; Ozawa, Y.; Ohkubo, K.; Fukuzumi, S. Structure and Spectroscopy of Oxyluciferin, the Light Emitter of the firefly bioluminescence. *J. Am. Chem. Soc.* **2009**, *131*, 11590–11605.
- (28) Suzuki, N.; Sato, M.; Okada, K.; Goto, T. Studies of firefly bioluminescence 0.1. Synthesis and spectral properties of firefly oxyluciferin; a possible emitting species in firefly bioluminescence. *Tetrahedron* **1972**, *28*, 4065–4074.
- (29) White, E. H.; Roswell, D. F. Analogs and derivatives of firefly oxyluciferin, the light emitter in firefly bioluminescence. *Photochem. Photobiol.* **1991**, *53*, 131–136.
- (30) Leont'eva, O. V.; Vlasova, T. N.; Ugarova, N. N. Dimethyl- and monomethyloxyluciferins as analogs of the product of the bioluminescence reaction catalyzed by firefly luciferase. *Biochemistry (Moscow)* **2006**, *71*, 51–55.
- (31) Vlasova, T. N.; Leontieva, O. V.; Ugarova, N. N. Interaction of dimethyl- and monomethyloxyluciferin with recombinant wild-type and mutant firefly luciferases. *Biochemistry (Moscow)* **2006**, *71*, 555–559.
- (32) Naumov, P.; Kochunnonny, M. Spectral-Structural Effects of the Keto-Enol-Enolate and Phenol-Phenolate Equilibria of Oxyluciferin. *J. Am. Chem. Soc.* **2010**, *132*, 11566–11579.
- (33) White, E. H.; Rapaport, E.; Hopkins, T. A.; Seliger, H. H. Chemi- and bioluminescence of firefly luciferin. *J. Am. Chem. Soc.* **1969**, *91*, 2178–2180.
- (34) White, E. H.; Steinmetz, M. G.; Miano, J. D.; Wildes, P. D.; Morland, R. Chemi-luminescence and bioluminescence of firefly luciferase. *J. Am. Chem. Soc.* **1980**, *102*, 3199–3208.
- (35) Jensen, F. In *Introduction to Computational Chemistry*; Wiley: New York, 1999; p 89.
- (36) Cramer, C. J. In *Essentials of Computational Chemistry. Theories and Models*; Wiley: New York, 2002; p 139.
- (37) Orlova, G.; Goddard, J. D.; Brovko, L. Y. Theoretical Study of the Amazing Firefly Bioluminescence: The Formation and Structures of the Light Emitters. *J. Am. Chem. Soc.* **2003**, *125*, 6962–6971.
- (38) Foresman, J. B.; Head-Gordon, J. A.; Frisch, M. J. Towards a systematic molecular-orbital theory for excited-states. *J. Phys. Chem.* **1992**, *96*, 135–49.
- (39) Becke, A. D. Density-Functional Thermochemistry 0.3. The role of Exact Exchange. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (40) Hehre, W. J.; Ditchfield, R.; Pople, J. A. Self-Consistent Molecular Orbital Methods. XII: Further Extensions of Gaussian-Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules. *J. Chem. Phys.* **1972**, *56*, 2257–2261.
- (41) Branchini, B. R.; Magyar, R. A.; Murtiashaw, M. H.; Anderson, S. M.; Zimmer, M. Site-directed mutagenesis of histidine 245 in firefly luciferase: A proposed model of the active site. *Biochemistry* **1998**, *37*, 15311–15319.
- (42) Sandalova, T. P.; Ugarova, N. N. Model of the active site of firefly luciferase. *Biochemistry (Moscow)* **1999**, *64*, 962–967.
- (43) Liu, Y.-J.; De Vico, L.; Lindh, R. *Ab initio* investigation on the chemical origin of the firefly bioluminescence. *J. Photochem. Photobiol. A* **2008**, *194*, 261–267.

- (44) Gross, E. K. U.; Kohn, W. Time-Dependent Density-Functional Theory. *Adv. Quantum Chem.* **1990**, *21*, 255–291.
- (45) Casida, M. E. In *Recent Advances in Density Functional Methods*; Chong, D. P., Ed.; World Scientific: Singapore, 1995; p 155.
- (46) Zerner, M. C. In *Reviews of Computational Chemistry*; Lipkowitz, K. B.; Boyd, D. B., Ed.; VCH: New York, 1991.
- (47) De Mello, P. C.; Hehenberger, M.; Zernert, M. C. Converging SCF calculations on excited states. *J. Quantum Chem.* **1982**, *21*, 251–258.
- (48) Burcl, R.; Amos, R. D.; Handy, N. C. Study of excited states of furan and pyrrole by time-dependent density functional theory. *Chem. Phys. Lett.* **2002**, *355*, 8–18.
- (49) Gross, L. A.; Baird, G. S.; Hoffman, R. C.; Baldrige, K. K.; Tsieng, R. Y. The structure of the chromophore within DsRed, a red fluorescent protein from coral. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 11990–11995.
- (50) Ugarova, N. N.; Brovko, L. Y. Protein structure and bioluminescent spectra for firefly bioluminescence. *Luminescence* **2002**, *17*, 321–330.
- (51) Ren, A.-M.; Goddard, J. D. Predictions of the electronic absorption and emission spectra of luciferin and oxyluciferins including solvation effects. *J. Photochem. Photobiol. B* **2005**, *81*, 163–170.
- (52) Foresman, J. B.; Keith, T. A.; Wiberg, K. B.; Snoonian, J.; Frisch, M. J. Solvent effects 0.5. Influence of cavity shape; truncation of electrostatics; and electron correlation ab initio reaction field calculations. *J. Phys. Chem.* **1996**, *100*, 16098–16104.
- (53) Yang, T.; Goddard, J. D. Predictions of the Geometries and Fluorescence Emission Energies of Oxyluciferin. *J. Phys. Chem. A* **2007**, *111*, 4489–4497.
- (54) Andersson, K.; Malmqvist, P. A.; Roos, B. O.; Sadlej, A. J.; Wolinski, K. Second-order perturbation theory with a CAS-SCF reference function. *J. Phys. Chem.* **1990**, *94*, 5483–5488.
- (55) Andersson, K.; Malmqvist, P. A.; Roos, B. O. Second-order perturbation theory with a complete active space self-consistent field reference function. *J. Chem. Phys.* **1992**, *96*, 1218–1226.
- (56) Hariharan, P. C.; Pople, J. A. The influence of polarization functions on molecular orbital hydrogenation. *Theor. Chim. Acta* **1973**, *28*, 213–222.
- (57) Nakatani, N.; Hasegawa, J.-Y.; Nakatsuji, H. Red Light in Chemiluminescence and Yellow-Green Light in Bioluminescence: Color-Tuning Mechanism of Firefly; *Photinus pyralis*: Studied by the Symmetry-Adapted Cluster-Configuration Interaction Method. *J. Am. Chem. Soc.* **2007**, *129*, 8756–8765.
- (58) Nakatsuji, H.; Hirao, K. Cluster expansion of the wavefunction. Symmetry-adapted-cluster expansion; its variational determination; and extension of open-shell orbital theory. *J. Chem. Phys.* **1978**, *68*, 2053–2066.
- (59) Nakatsuji, H. In *Computational Chemistry; Reviews of current Trends*; Leszczynski, J., Ed.; World Scientific: Singapore, 1996; p 62.
- (60) Dreuw, A.; Weisman, J. L.; Head-Gordon, M. Long-range charge-transfer excited states in time-dependent density functional theory require non-local exchange. *J. Chem. Phys.* **2003**, *119*, 2943–2947.
- (61) Fujimoto, K.; Hayashi, S.; Hasegawa, J.-Y.; Nakatsuji, H. Theoretical Studies on the Color-Tuning Mechanism in Retinal Proteins. *J. Chem. Theory Comput.* **2007**, *3*, 605–618.
- (62) Barone, V.; Cossi, M.; Tomasi, J. A new definition of cavities for the computation of solvation free energies by the polarizable continuum model. *J. Chem. Phys.* **1997**, *107*, 3210–3222.
- (63) Li, Z.-W.; Ren, A.-M.; Guo, J.-F.; Yang, T.; Goddard, J. D.; Feng, J.-K. Color-Tuning Mechanism in Firefly Luminescence: Theoretical Studies on Fluorescence of Oxyluciferin in Aqueous Solution Using Time Dependent Density Functional Theory. *J. Phys. Chem. A* **2008**, *112*, 9796–9800.
- (64) Min, C.-G.; Ren, A.-M.; Guo, J.-F.; Li, Z.-W.; Zou, L.-Y.; Goddard, J. D.; Feng, J.-K. A Time-Dependent Density Functional Theory Investigation on the Origin of Red Chemiluminescence. *ChemPhysChem* **2010**, *11*, 251–259.
- (65) Cohen, A. J.; Mori-Sánchez; Yang, W. Insights into current limitations of density functional theory. *Science* **2008**, *321*, 792–794.
- (66) Perdew, J. P.; Wang, Y. Accurate and simple analytic representation of the electron-gas correlation-energy. *Phys. Rev. B* **1992**, *45*, 13244–13249.
- (67) Klamt, A.; Schuurmann, G. COSMO - A new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc., Perkin Trans 2* **1993**, *5*, 799–805.
- (68) Li, Z.-W.; Min, C.-G.; Ren, A.-M.; Guo, J.-F.; Goddard, J. D.; Feng, J.-K.; Zuo, L. Theoretical Study of the Relationships between Excited State Geometry Changes and Emission Energies of Oxyluciferin. *Bull. Korean Chem. Soc.* **2010**, *31*, 895–900.
- (69) Conti, E.; Franks, N. P.; Brick, P. Crystal structure of firefly luciferase throws light on a superfamily of adenylate-forming enzymes. *Structure* **1996**, *4*, 287–298.
- (70) Nakatani, N.; Hasegawa, J.; Nakatsuji, H. Artificial color tuning of firefly luminescence, Theoretical mutation by tuning electrostatic interactions between protein and luciferin. *Chem. Phys. Lett.* **2009**, *469*, 191–194.
- (71) Tagami, A.; Ishibashi, N.; Kato, D.-I.; Taguchi, N.; Mochizuki, Y.; Watanabe, H.; Ito, M.; Tanaka, S. *Ab initio* quantum-chemical study on emission spectra of bioluminescent luciferases by fragment molecular orbital method. *Chem. Phys. Lett.* **2009**, *472*, 118–123.
- (72) Mochizuki, Y.; Tanaka, K.; Yamashita, K.; Ishikawa, T.; Nakano, T.; Amari, S.; Segawa, K.; Murase, T.; Tokiwa, H.; Sakurai, M. Parallelized integral-direct CIS(D) calculations with multilayer fragment molecular orbital scheme. *Theor. Chem. Acc.* **2007**, *117*, 541–553.
- (73) Head-Gordon, M.; Rico, R. J.; Oumi, M.; Lee, T. J. A doubles correction to electronic excited states from configuration interaction in the space of single substitutions. *Chem. Phys. Lett.* **1994**, *219*, 21–29.
- (74) Navizet, I.; Liu, Y.-J.; Xiao, H. Y.; Fang, W. H.; Lindh, R. Color-tuning mechanism of firefly investigated by multi-configurational perturbation method. *J. Am. Chem. Soc.* **2010**, *132*, 706–712.
- (75) Milne, B. F.; Marques, N. A.; Nogueira, F. Fragment molecular orbital investigation of the role of AMP protonation in firefly luciferase pH-sensitivity. *Phys. Chem. Chem. Phys.* **2010**, *12*, 14285–14293.
- (76) Dawson, R. M. C.; Elliott, D. C.; Elliott, W. H.; Jones, K. M. *Data for Biochemical Research*; Oxford University Press: Oxford, England, 1989.
- (77) Min, C.-G.; Ren, A.-M.; Guo, J.-F.; Goddard, J. D.; Sun, C. C. Theoretical investigation on the origin of yellow-green firefly bioluminescence by time-dependent density functional theory. *ChemPhysChem* **2010**, *11*, 2199–2204.
- (78) Siegbahn, P. E. M.; Eriksson, L.; Himo, F.; Pavlov, M. Hydrogen atom transfer in ribonucleotide reductase (RNR). *J. Phys. Chem. B* **1998**, *102*, 10622–10629.
- (79) Siegbahn, P. J. Theoretical study of the substrate mechanism of ribonucleotide reductase. *J. Am. Chem. Soc.* **1998**, *120*, 8417–8429.
- (80) Blomberg, M. R. A.; Siegbahn, P. E. M.; Babcock, G. T. Modeling electron transfer in biochemistry: A quantum chemical study of charge separation in *Rhodobacter sphaeroides* and photosystem II. *J. Am. Chem. Soc.* **1998**, *120*, 8812–8824.
- (81) Fernandes, P. A.; Ramos, M. J. Theoretical Studies on the Mechanism of Inhibition of Ribonucleotide Reductase by (E)-2'-Fluoromethylene-2'-deoxycytidine-5'-diphosphate. *J. Am. Chem. Chem. Soc.* **2003**, *125*, 6311–6322.
- (82) Mao, Y. Dynamics studies of luciferase using elastic network model: how the sequence distribution of luciferase determines its color. *Protein Eng., Des. Sel.* **2010**, DOI: 10.1093/protein/gzq109.
- (83) Tirion, M. M. Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Phys. Rev. Lett.* **1996**, *77*, 1905–1908.
- (84) Cai, D.; Marques, M. A.; Nogueira, F. Accurate Color Tuning of Firefly Chromophore by Modulation of Local Polarization Electrostatic Fields. *J. Phys. Chem. B* **2010**, *115*, 329–332.
- (85) Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **1964**, *136*, B864–B871.
- (86) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140*, A1133–A1138.

(87) Chen, S.-F.; Liu, Y.-J.; Navizet, I.; Ferré, N.; Fang, W.-H.; Lindh, R. Systematic Theoretical Investigation on the Light Emitter of Firefly. *J. Chem. Theory Comput.* **2011**, DOI: 10.1021/ct200045q.

(88) Pinto da Silva, L.; Esteves da Silva, J. C. G. Computational Investigation of the effect of pH on the Color of Firefly Bioluminescence by DFT. *ChemPhysChem* **2011**, DOI: 10.1002/cphc.201000980.

(89) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

Molecular Dynamics Simulations Predict a Favorable and Unique Mode of Interaction between Lithium (Li^+) Ions and Hydrophobic Molecules in Aqueous Solution

Andrew S. Thomas and Adrian H. Elcock*

Department of Biochemistry, University of Iowa, Iowa City, Iowa 52242, United States

S Supporting Information

ABSTRACT: We report the results of molecular dynamics simulations of the interaction thermodynamics of group I cations with hydrophobic molecules in water using a variety of nonpolarizable force fields. Surprisingly, we find that the Li^+ ion is predicted to form thermodynamically favorable interactions with methane and neopentane using a new mode of recognition that is intermediate between a direct contact and a solvent-separated complex. Further simulations show that this favorable interaction is only predicted by ion parameter sets that correctly reproduce Li^+ 's experimental hydration number.

INTRODUCTION

While the behavior of ions in aqueous solution has long been subjected to study through both theory and experiment,¹ only recently have force field parameters for ions reached a level of maturity at which molecular simulation methods can reproduce simultaneously both the thermodynamic and structural features of the ion–water interaction.² The development of accurate parameter sets for the group I cations, in particular, opens the way to a detailed study of their interactions with other biomolecular solutes in aqueous solution. As one step in this direction, we have used a series of long molecular dynamics (MD) simulations to undertake a comparative study of the interaction thermodynamics of three group I cations, Li^+ , Na^+ , and K^+ , with hydrophobic molecules in water.

The few previous computational studies examining interactions between ions and hydrophobic molecules have suggested that their nature can depend significantly on the size of the ion. Using a simplified 2D water model, Hribar et al.³ showed that in an aqueous solution containing a small, charge-dense cation (reminiscent of Li^+) an inserted hydrophobic solute preferred to adopt a solvent-separated configuration relative to the ion; in contrast, when a larger, charge-diffuse cation (similar to Cs^+) was present in the solution, the hydrophobic solute was observed to penetrate the hydration shell of the ion and “bind” to the ion's surface. Following that work, others have used more conventional molecular dynamics (MD) simulations to model interactions between ions and hydrophobic groups. Shinto et al.⁴ explored how the ions Na^+ , Cl^- , and tetramethylammonium, $(\text{CH}_3)_4\text{N}^+$, interact with a united-atom model of methane. In the case of the comparatively small, charge-dense ion, Na^+ , a thermodynamically favorable solvent-separated interaction was observed, but closer interaction, in the form of a direct contact between the ion and the methane, was apparently prevented due (presumably) to the energetically unfavorable desolvation that would be incurred by the ion. In the case of the much larger, charge-diffuse tetramethylammonium ion, however, direct binding to the methane was observed, in a manner qualitatively

similar to the conventional association of two hydrophobic molecules. Similar observations were reported by Lund et al.⁵ for interactions of monovalent anions (F^- and I^-) with a nanometer-sized hydrophobic solute; again, the charge-dense F^- ion was effectively excluded from the nonpolar surface, while the charge-diffuse I^- ion bound quite readily. Taken together therefore, all of these previous works suggest a simple, general rule for ion–hydrophobe interactions: large, charge-diffuse ions have the potential to bind directly to hydrophobic groups, but small, charge-dense ions do not (although they may adopt favorable solvent-separated configurations).^{3–5}

The molecular simulations reported here provide evidence that this line of thinking may not always be appropriate: the simulations predict a surprising, thermodynamically favorable interaction between the small, charge-dense Li^+ ion and hydrophobic molecules in aqueous solution. We further show that this favorable interaction is a robust prediction of all tested force fields that correctly reproduce the hydration number of the Li^+ ion.

METHODS

Free energies of interaction between ions and hydrophobic molecules were obtained from explicit solvent MD simulations performed with the GROMACS 4.0 software.⁶ In all simulations, a single hydrophobic solute (methane or neopentane) and a single ion were placed in a $25 \times 25 \times 25$ Å simulation box containing approximately 500 explicitly modeled water molecules such that the overall density was 1 g/cm^3 . For each type of ion, independent simulations were performed with all three of the ion parameter sets recently derived by Joung and Cheatham,² these three sets have been developed for use, respectively, with the SPC/E,⁷ TIP3P,⁸ and TIP4P-Ew⁹ water models. In the case of the Li^+ ion, simulations were also performed with Åqvist's ion parameters¹⁰ in combination with SPC/E, SPC,¹¹ and TIP3P

Received: September 14, 2010

Published: March 03, 2011

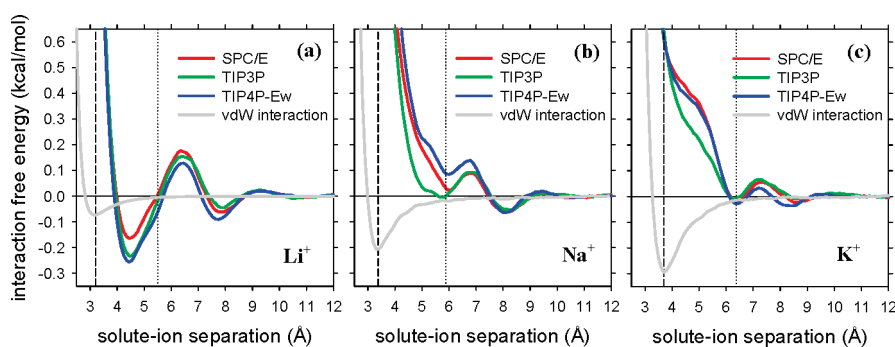


Figure 1. Computed interaction free energies for the association of methane with (a) Li^+ , (b) Na^+ , and (c) K^+ as a function of the ion–methane distance. The red, green, and blue lines show results obtained with Joung and Cheatham’s ion parameters specifically derived for use with the water models identified in the inset. The solid gray line in each figure is the direct interaction between the ion and the hydrophobic solute calculated in vacuum. The positions of the van der Waals contact minimum (CM) and the solvent-separated minimum (SSM) are indicated by dashed and dotted vertical lines, respectively. The expected position of a SSM for Li^+ is shown as 2.34 Å further than the CM position; this is based on a linear fit between the positions of the SSM and the first peak in the cation–water radial distribution function for the three group I cations. Error bars for each data set are not shown for purposes of clarity; however, they amount to less than 0.05 kcal/mol.

water, and with Jensen and Jorgensen’s ion parameters¹² in TIP4P water.⁸ The hydrophobic solutes methane and neopentane were both described using parameters from the OPLS all-atom force field;¹³ additional simulations of the Li^+ –methane interaction were also performed with methane described by the OPLS united-atom force field.¹⁴ To be consistent with the parametrization of the respective ions, the Lorentz-Berthelot² or geometric^{10,12} mixing rules were used for all van der Waals interactions that involved ions; to be consistent with the OPLS parametrization scheme,¹³ geometric mixing rules were used to describe van der Waals interactions between the hydrophobic solutes and water.

All MD simulations used a standard protocol. Short-range non-bonded interactions were directly calculated up to a 10 Å cutoff; long-range electrostatic interactions beyond this cutoff were calculated using the particle mesh Ewald (PME) method.¹⁵ The temperature was maintained at 298 K using the Nosé–Hoover thermostat,¹⁶ and pressure was maintained at 1 atm using the Parrinello–Rahman barostat.¹⁷ Covalent bonds in the methane were constrained using the LINCS algorithm,¹⁸ allowing for a 2 fs time step to be used. Simulations were first heated up to 298 K over the course of 1 ns, then equilibrated for a further 10 ns, before “production” simulations of 500 ns were conducted. As we have shown previously,¹⁹ these long simulation times allow for converged estimates of the association thermodynamics to be calculated directly from completely unbiased MD simulations. In the present study, the free energy of interaction, ΔG , as a function of the ion–hydrophobe distance was determined from the calculated radial distribution function, $g(r)$, of the central carbon-to-ion distance using $\Delta G = -RT \ln g(r)$. Radial distribution functions were calculated using the g_rdf utility of GROMACS.⁶

Since the results obtained from the above simulations turned out to be so surprising in the case of Li^+ (see the Results), a large number of additional control simulations of the Li^+ –methane system were performed to establish the robustness of the predicted behavior. To rule out an influence from the choice of mixing rules used to describe the ion–hydrophobe van der Waals interactions, additional simulations were performed using alternative mixing rules (see the Supporting Information for details). To explore the role of the van der Waals parameters assigned to the methane, additional simulations were performed using AMBER99 van der Waals parameters²⁰ (see Figure S1, Supporting Information). Finally, at the behest of a reviewer, additional simulations were performed with a larger, 35 Å simulation box

(Figure S2, Supporting Information). As shown in the Supporting Information, all of these control simulations produced results qualitatively identical with the results reported herein.

In order to explore the sensitivity of the observed behavior of Li^+ to the presence of additional salt, MD simulations of the Li^+ –methane interaction were also performed in a 1 M LiCl solution; these simulations involved the addition of eight Li^+ and nine Cl^- ions to the system but were run in an otherwise identical manner to those conducted in pure water. Finally, to explore whether results similar to those obtained with the Li^+ –methane system might also be found in some other systems MD simulations were used to study the interaction thermodynamics of methane with the NH_4^+ cation,¹³ and with the F^- anion;² independent simulations of the latter interactions were performed using all three of the Joung and Cheatham parameter sets.

RESULTS AND DISCUSSION

The interaction free energies computed for methane’s interaction with Li^+ , Na^+ , and K^+ in water are plotted as a function of the ion–hydrophobe separation distance in Figure 1a, b, and c, respectively. The colored lines in each panel show the results computed with the three water model-dependent ion parameter sets developed by Joung and Cheatham;² clearly, all three parameter sets give very similar results (corresponding results obtained with other parameter sets are considered later in this manuscript). The most surprising and interesting results of the simulations reported here are to be found in the plots of the Li^+ –methane interaction free energy versus the distance (Figure 1a). Two aspects of these plots are particularly notable. First, the Li^+ –methane interaction is predicted to be thermodynamically favorable at distances between 4.0 and 5.5 Å, with a strength approximately one-third that of the more conventional hydrophobic interaction between two methane molecules.²¹ This observation of a thermodynamically favorable interaction between Li^+ and methane is a major surprise since, as noted in the Introduction, the conventional view of ion–hydrophobe interactions is that they should be thermodynamically unfavorable, owing to the cost of disrupting the ion’s hydration shell.³ Certainly, this conventional thinking is borne out for methane’s interaction with both Na^+ (Figure 1b) and K^+ (Figure 1c): neither ion shows a tendency to form a short-range favorable

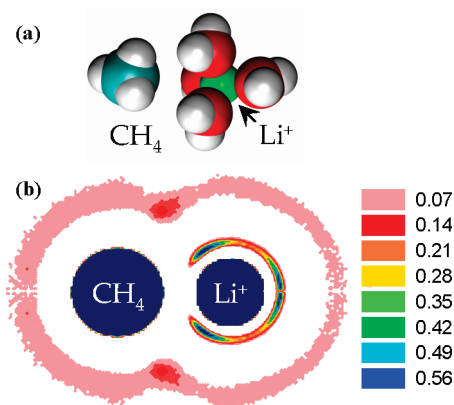


Figure 2. (a) MD snapshot of the Li⁺–methane complex at its free energy minimum configuration. (b) 2-D projection of water density around the free energy minimum configuration (molecules/Å³). This plot was computed using all snapshots in which the Li⁺–methane separation was 4.5–4.6 Å in the MD simulation that used the Joung and Cheatham SPC/E parameters. Essentially identical results were obtained using Joung and Cheatham’s TIP3P and TIP4P-Ew parameters.

interaction with the hydrophobic molecule. But since Li⁺ is the most charge-dense of the group I cations,¹ it would be expected to be even more adversely affected by the approach of a nonpolar molecule than either Na⁺ or K⁺. The simulation results shown in Figure 1a, however, clearly challenge this notion.

The second notable aspect of the results obtained for the Li⁺–methane interaction (Figure 1a) is that the distance at which the interaction is most favorable (~4.5 Å) corresponds to neither of the two usual modes of molecular interaction: it is too long to be a direct van der Waals contact interaction (indicated by the dotted gray vertical line) and too short to be a solvent-separated interaction of the type sometimes observed with other ions (the expected position of this minimum is indicated by the dotted gray vertical line). Instead, an examination of “snapshots” taken from the simulations shows that at this separation distance the Li⁺ ion retains a tetrahedrally coordinated hydration shell¹ by approaching the hydrophobic molecule along a line trisecting three of the Li⁺–H₂O “bonds” (Figure 2a). As a result, the complex encloses a ~1.5-Å-long region of the vacuum separating the ion from its hydrophobic binding partner.

A more quantitative illustration of the water distribution around the Li⁺–methane complex can be obtained from the two-dimensional projection of the water oxygen density shown in Figure 2b. This plot clearly hints at the retention of the highly structured first hydration shell around the Li⁺ ion and also shows, interestingly, that water density is increased significantly (by a factor of ~2) in the region where the methane’s first hydration shell intersects with the diffuse second hydration shell of the Li⁺ ion. Aside from this latter effect, however, the Li⁺ ion does not appear to induce significant additional structuring of the water surrounding the hydrophobic molecule. The retention of the tetrahedral hydration waters, even in complex with methane, suggests therefore that Li⁺ acts as a “permanently” hydrated ion [Li–(H₂O)₄]⁺ in solution, similar to some small divalent and trivalent ions.²²

Although it is often difficult to distinguish cause from effect when analyzing simulated behavior, we have attempted to identify the factors that drive the formation of the favorable Li⁺–methane contact by decomposing the interaction thermodynamics into individual energetic components. To this end, we have taken 5 million structural snapshots from each simulation and extracted the

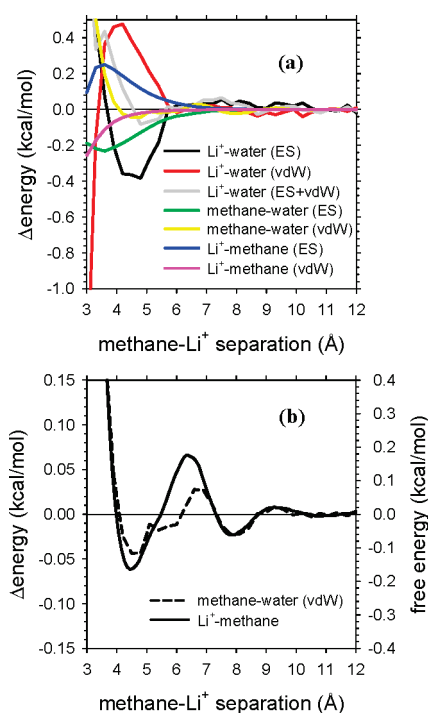


Figure 3. (a) Energetic contributions to the Li⁺–methane interaction plotted as a function of the Li⁺–methane distance. Lines shown correspond to data from simulations that used the Joung and Cheatham SPC/E parameter set. (b) The change in the methane–water van der Waals interaction energy as a function of the Li⁺–methane distance is plotted as a dashed line against the left-hand y axis (these data are replotted from part a); for comparison, the interaction free energy as a function of the Li⁺–methane distance is plotted as a solid line against the right-hand y axis (these data are replotted from Figure 1a).

electrostatic and van der Waals components of each of the following interactions: (a) Li⁺–water, (b) Li⁺–methane, and (c) methane–water; the results of this analysis for the SPC/E set of Joung and Cheatham parameters are shown in Figure 3 (corresponding results for the other Joung and Cheatham parameter sets are shown in Figure S3, Supporting Information). Interestingly, the Li⁺–water electrostatic interaction energy becomes significantly more favorable as the Li⁺ and methane approach one another and reaches its minimum value (stabilized by –0.4 kcal/mol) when the Li⁺–methane distance reaches its free energy minimum distance (4.5 Å). This favorable change in energy is opposed by a corresponding unfavorable change in the Li⁺–water van der Waals interaction (compare red and black lines in Figure 3a), but the net effect (i.e., the sum of the electrostatic and van der Waals terms) is a slightly favorable contribution to the interaction, albeit one that reaches a minimum value at a somewhat farther distance (~5 Å) than the free energy minimum distance. Similar effects, but differing markedly in magnitude, are seen with the other Joung and Cheatham parameter sets (see Figures S3a and b, Supporting Information). The methane–water electrostatic interaction (green line) also becomes more energetically favorable as the Li⁺ and methane approach one another and is most favorable when the two are in direct contact with one another; this component, however, appears to be effectively canceled by a corresponding unfavorable change in the Li⁺–methane electrostatic interaction (blue line). As expected, the Li⁺–methane van der Waals interaction (pink line) becomes increasingly favorable as the two molecules approach, but its contribution is very small at a separation distance of 4.5 Å (–0.03 kcal/mol) and is not likely,

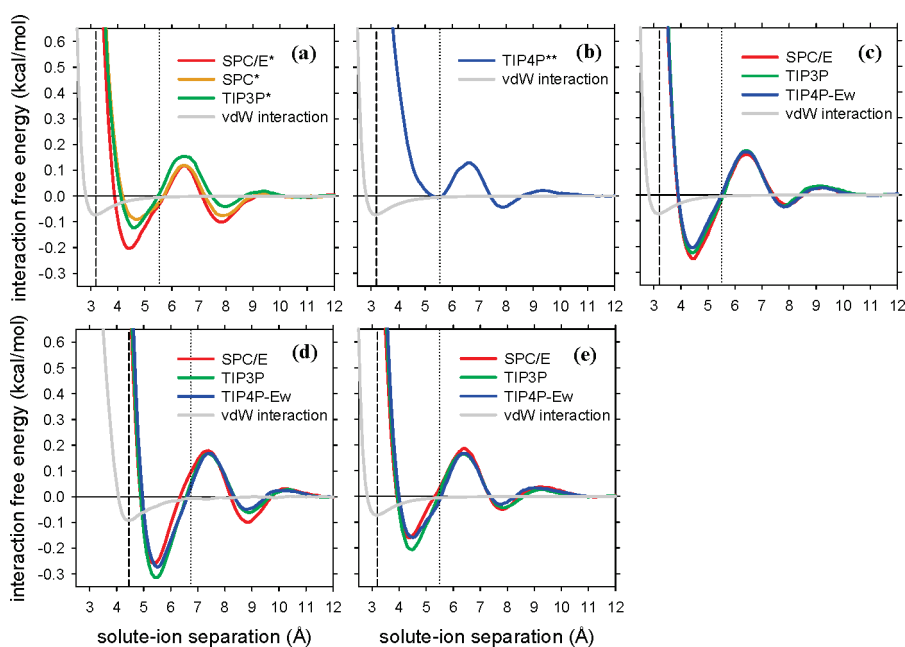


Figure 4. Computed interaction free energies for the association of Li^+ with (a) methane using Åqvist's (denoted *) ion parameters, (b) methane using Jensen and Jorgensen's (denoted **) ion parameters, (c) united-atom methane using Joung and Cheatham's ion parameters, (d) neopentane using Joung and Cheatham's ion parameters, and (e) methane in 1 M LiCl solution using Joung and Cheatham's ion parameters.

therefore, to be a significant determinant of the favorable Li^+ –methane interaction. The final energetic contribution to consider is the methane–water van der Waals interaction (yellow line in Figure 3a). Intriguingly, the shape of this curve follows the free energy curve very nicely with all three of the Joung and Cheatham parameter sets (see Figure 3b and Figure S3c and d, Supporting Information), which suggests that this interaction is a potential determinant of the observed interaction thermodynamics. A more complete analysis of the underlying thermodynamics would, of course, require an analysis of entropic contributions (see e.g. refs 23 and 24), but that is beyond the scope of the present work.

Since the prediction of a thermodynamically favorable Li^+ –methane interaction is a major surprise and is likely to be controversial, we have been careful to explore whether its prediction is robust to changes in the simulation parameters. Figure 4a shows the results computed using parameters for Li^+ derived by Åqvist; since his ion parameter set has been used in conjunction with a number of three-site water models in the literature, we performed independent simulations with each of the TIP3P, SPC, and SPC/E water models. With all three models, we again obtain a thermodynamically favorable interaction at a separation distance of ~ 4.5 Å (Figure 4a). Interestingly, however, the interaction free energy is computed to be significantly more favorable with the SPC/E model than with either of the other two water models; a likely explanation for this result is outlined below.

Figure 4b shows the results computed using Li^+ parameters derived by Jensen and Jorgensen.¹² In this case, it is clear that a thermodynamically favorable Li^+ –methane interaction is not predicted. At first sight this result appears to call into question the significance of the results reported earlier. Below, however, we present evidence that this particular parameter set may be compromised by having been derived to fit experimental data that have since been subject to revision.

Figure 4c shows the results computed using Joung and Cheatham's Li^+ parameters in combination with an OPLS

united-atom description for the methane.¹⁴ In this case, results essentially identical to those obtained with the all-atom methane model are obtained, indicating that the observed interaction is not dependent on the details of the methane's modeling. The results of additional simulations, which explored (a) changing the mixing rules used to describe van der Waals interactions and (b) increasing the simulation box size, are shown in Figures S1 and S2, Supporting Information, respectively; neither change resulted in any significant change to the computed thermodynamics of the Li^+ –methane interaction.

Figure 4d shows the results computed using Joung and Cheatham's ion parameters in simulations of the interaction of Li^+ with the larger hydrophobic molecule neopentane. From this, it is clear that the prediction of a thermodynamically favorable interaction between the Li^+ ion and a hydrophobic molecule is not restricted to methane; the Li^+ –neopentane interaction free energy is, in fact, predicted to be somewhat favorable (-0.25 to -0.30 kcal/mol).

Finally, Figure 4e shows the results computed using Joung and Cheatham's ion parameters in simulations of the Li^+ –methane interaction in a 1 M LiCl solution. Again, a favorable Li^+ –methane interaction is retained: even competition from halide ions, therefore, appears unable to completely suppress the ion–hydrophobe interaction.

Focusing again on those panels of Figures 1 and 4 that plot the computed free energies of the Li^+ –methane interaction in pure water, we can make the following statements: (a) The most favorable Li^+ –methane interactions are predicted by the three Joung and Cheatham parameter sets and by Åqvist's parameters when used with the SPC/E water model. (b) Somewhat weaker but still favorable Li^+ –methane interactions are predicted by Åqvist's parameters when used with the SPC and TIP3P water models. (c) A favorable Li^+ –methane interaction is not predicted by the Jensen and Jorgensen parameters. Since these are significant discrepancies, it is important both to try to identify

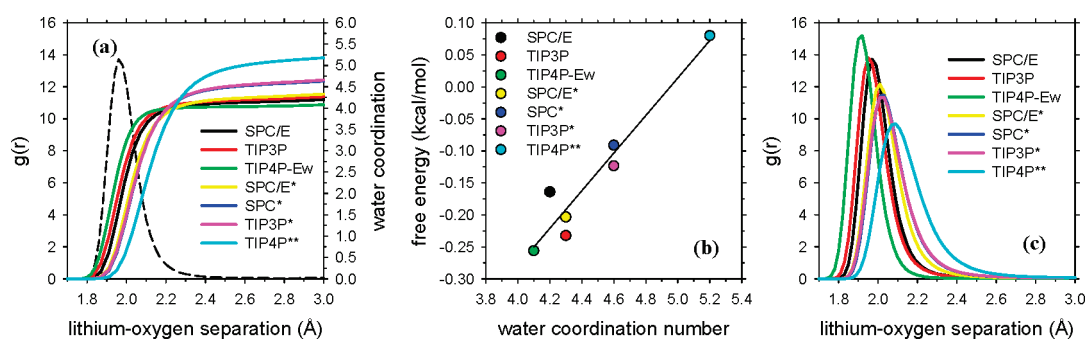


Figure 5. Effects of ion and water model parameters on the Li^+ hydration structure. (a) Li^+ –water coordination numbers (plotted against the right-hand y axis) and the Li^+ –water oxygen radial distribution function in SPC/E (left-hand y axis, dashed line), both plotted as a function of the Li^+ –water oxygen distance. (b) Correlation between the free energy at the Li^+ –methane free energy minimum (at a separation of 4.55 Å) and the Li^+ –water coordination number computed with the same parameter set. (c) Li^+ –water oxygen radial distribution functions plotted as a function of the Li^+ –water oxygen distance. The water model-dependent ion parameters of Joung and Cheatham, Åqvist (denoted *), and Jensen and Jorgensen (denoted **) were used.

their origins and to determine which of the predicted results is likely to be the most realistic.

In the absence of direct experimental data on the thermodynamics of the Li^+ –methane interaction, we must attempt to assess the quality of the various parameter sets by examining their abilities to reproduce alternative sources of experimental data. One obvious measure to consider is the hydration free energy of the Li^+ ion since this is likely to be a major determinant of its observed interaction thermodynamics with methane in aqueous solution. But both the Joung and Cheatham, and Jensen and Jorgensen parameter sets have been explicitly parametrized to reproduce Marcus' estimate of Li^+ 's hydration free energy of -113 kcal/mol,²⁵ effectively ruling this out as a possible cause of the differences observed between the two parameter sets.

A second measure that can be used to assess the quality of the various parameter sets is the water coordination number of the isolated solutes. All of the various simulation models used here predict the number of waters in the very diffuse hydration shell around methane to average ~ 19 – 21 (Figure S4, Supporting Information). Encouragingly, these estimates are consistent both with previous computational studies²⁶ and, more importantly, with a recent experimental estimate of 20.²⁷

But the various simulation models differ significantly in the water coordination numbers that they predict for the isolated Li^+ ion (see Figure 5a). The mean coordination numbers that we obtain from our simulations are, for the Joung and Cheatham Li^+ parameters: 4.2, 4.3, and 4.1 using the SPC/E, TIP3P, and TIP4P-Ew models, respectively; for the Åqvist Li^+ parameters: 4.3, 4.6, and 4.6 using the SPC/E, SPC, and TIP3P models, respectively; and for the Jensen and Jorgensen Li^+ parameters (in TIP4P): 5.2. All of these numbers agree well with those reported in previously published simulation studies that used the same force fields.^{12,28} More interestingly, however, there is a clear correlation between the computed hydration numbers of the isolated Li^+ ion and the predicted free energy of the Li^+ –methane interaction obtained with the same parameter set (see Figure 5b): those parameter sets that produce lower hydration numbers predict significantly more favorable interaction free energies for the Li^+ –methane complex.

This trend becomes significant when we note that recent experimental studies indicate that the water coordination number of the Li^+ ion is 4,^{1,29} and that similar numbers have been obtained from polarizable MD studies,³⁰ from full quantum mechanical calculations of ion–water clusters,³¹ and from QM-MM calculations of ions in aqueous solution.³² On the basis of these previous studies, therefore,

it appears that the best structural description of Li^+ 's hydration is provided by the three Li^+ parameter sets derived by Joung and Cheatham and by Åqvist's parameters when used with the SPC/E water model. All things being equal, therefore, we anticipate that the Li^+ –methane interaction thermodynamics predicted by the Joung and Cheatham parameter sets should be the most realistic, and that the possibility of the predicted favorable Li^+ –methane interaction being real is therefore significant.

The results obtained with Jensen and Jorgensen's Li^+ parameters suggest that the presence of an additional water molecule in the ion's hydration shell is sufficient to prevent the formation of a thermodynamically favorable Li^+ –methane complex. This is consistent with the clearly tetrahedral arrangement of water molecules around the Li^+ ion in the MD snapshots of the Li^+ –methane complex (Figure 2). But what is the origin of the apparently excessive hydration number obtained with the Jensen and Jorgensen parameters? It appears to be the position of the first peak in the Li^+ –water oxygen RDF, which was explicitly used—together with the ion's experimental hydration free energy—as a target of the parametrization process.¹² At the time that Jensen and Jorgensen's work was conducted, the generally accepted position of this peak—which was a weighted average of disparate experimental measurements—was at a distance of 2.08 Å for Li^+ in a dilute solution,³³ but this value has since been revised downward by 0.12 Å to ~ 1.94 – 1.98 Å on the basis of more recent experimental measurements.¹ Accordingly, a comparison of the Li^+ –water oxygen RDFs computed with the various parameter sets used here shows that the Jensen and Jorgensen parameters produce a peak that is clearly shifted to a further distance than those obtained with other parameter sets (Figure 5c). Interestingly, the next furthest shifted peaks are those resulting from the use of Åqvist's parameters with the SPC and TIP3P models; these parameters, it will be recalled, predicted a noticeably weaker free energy for the Li^+ –methane interaction than was obtained with the SPC/E model (Figure 4a). There appears, therefore, to be a clear relationship between (a) the position of the first peak in the Li^+ –water oxygen RDF, (b) the (integrated) water coordination number, and (c) the computed thermodynamics of the Li^+ –methane interaction.

If we are correct in suggesting that the Jensen and Jorgensen Li^+ parameters may have been compromised by the subsequent change in the experimentally accepted position of the Li^+ –water oxygen RDF, it highlights an unavoidable danger faced by all simulation practitioners involved in the parametrization of force fields: even the most carefully derived parameter sets can be

brought into question if the source experimental data are later revised. In passing, we note that for Na^+ and K^+ we observe no qualitative differences between the predictions of the Jensen and Jorgensen parameters and those of other tested parameter sets (data not shown).

Of course, an important caveat to the above discussion is that we still cannot state with any certainty that the thermodynamically favorable Li^+ –methane interaction predicted in Figure 1a is correct; it is quite possible, for example, that despite the apparently incorrect hydration number obtained with the Jensen and Jorgensen parameters that its prediction of an unfavorable interaction might ultimately prove to be realistic. Given that the highly charge-dense nature of the Li^+ ion is likely to induce some degree of electronic polarization in nearby water molecules, it may be that a resolution of this uncertainty might be forthcoming from the use of one or more of the polarizable force fields which are actively being developed in a number of laboratories;^{30,34} it is also possible, however, that a quantum mechanical treatment might be required.^{31,32}

CONCLUSIONS

Despite the above caveat, we can state clearly that those parameter sets that currently appear to be more realistic—at least in terms of their structural description of Li^+ 's interaction with water—predict a more thermodynamically favorable interaction between Li^+ and methane. We can further state that an examination of the structure of the favorable Li^+ –methane complex shows it to represent a fundamentally new mode of molecular interaction that appears halfway between a van der Waals contact and a solvent-separated interaction. We can also assert, from the results shown in Figure 1b and c, that similar kinds of favorable interaction are not obtained with either Na^+ or K^+ and, on the basis of additional simulations reported in the Supporting Information, that a favorable interaction is also not predicted for methane's interaction with the most charge-dense of the group VII anions, F^- , or with the NH_4^+ ion (Figures S5 and S6, respectively, Supporting Information). Given that simulation results reported by others also show no evidence of a favorable interaction between methane and Cl^- or $(\text{CH}_3)_4\text{N}^+$,⁴ the predicted formation of a favorable short-range interaction between a monovalent ion and a hydrophobic molecule in aqueous solution appears to be unique to the Li^+ ion.

In closing, we note that the unusual interaction observed here between Li^+ and hydrophobic molecules provides an intriguing potential explanation for the finding that Li^+ salts can act to denature a protein even when the corresponding Na^+ and K^+ salts serve as stabilizers.³⁵ In addition, the same interaction may well be an important factor in determining Li^+ 's anomalously low salting-out ability.^{19b,36} Finally, it provides yet another example of the subtle and sometimes highly surprising ways in which hydration-shell water molecules can mediate—both structurally and thermodynamically—the interactions of ions³⁷ or biomolecular solutes³⁸ in aqueous solution.

ASSOCIATED CONTENT

S Supporting Information. Interaction free energies for Li^+ with methane using altered van der Waals parameters, mixing rules, and box size dependence; energy decompositions of Li^+ –methane simulation snapshots; water coordination numbers around methane; and interaction free energies for F^- with methane and for NH_4^+ with methane. This information is available free of charge via the Internet at <http://pubs.acs.org>

AUTHOR INFORMATION

Corresponding Author

*E-mail: adrian-elcock@uiowa.edu.

ACKNOWLEDGMENT

Supported by NSF CAREER award 0448029 (A.H.E.).

REFERENCES

- (1) Marcus, Y. *Chem. Rev.* **2009**, *109*, 1346–1370.
- (2) Joung, I. J.; Cheatham, T. E. *J. Phys. Chem. B* **2008**, *112*, 9020–9041.
- (3) Hribar, B.; Southall, N. T.; Vlachy, V.; Dill, K. A. *J. Am. Chem. Soc.* **2002**, *124*, 12302–12311.
- (4) Shinto, H.; Morisada, S.; Higashitani, K. *J. Chem. Eng. Jpn.* **2005**, *38*, 465–477.
- (5) Lund, M.; Vácha, R.; Jungwirth, P. *Langmuir* **2008**, *24*, 3387–3391.
- (6) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *92*, 435–447.
- (7) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269–6271.
- (8) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (9) Horn, H. W.; Swope, W. C.; Pitner, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. *J. Chem. Phys.* **2004**, *120*, 9665–9678.
- (10) Åqvist, J. *J. Phys. Chem.* **1990**, *94*, 8021–8024.
- (11) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. In *Interaction Models for Water in Relation to Protein Hydration*; Pullman, B., Ed.; Reidel Publishing Company: Dordrecht, The Netherlands, 1981; p 311.
- (12) Jensen, K. P.; Jorgensen, W. L. *J. Chem. Theory Comput.* **2006**, *2*, 1499–1509.
- (13) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.
- (14) Jorgensen, W. L.; Madura, J. D.; Swenson, C. J. *J. Am. Chem. Soc.* **1984**, *106*, 6638–6646.
- (15) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pederson, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593.
- (16) (a) Nosé, S. *J. Chem. Phys.* **1984**, *81*, 511–519. (b) Hoover, W. G. *Phys. Rev.* **1985**, *31*, 1695–1697.
- (17) Parrinello, M.; Rahman, A. *J. Appl. Phys.* **1981**, *52*, 7182–7190.
- (18) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J. Comput. Chem.* **1997**, *18*, 1463–1472.
- (19) (a) Thomas, A. S.; Elcock, A. H. *J. Am. Chem. Soc.* **2006**, *128*, 7796–7806. (b) Thomas, A. S.; Elcock, A. H. *J. Am. Chem. Soc.* **2007**, *129*, 14887–14898. (c) Zhu, S.; Elcock, A. H. *J. Chem. Theory Comput.* **2010**, *6*, 1293–1306. (d) Thomas, A. S.; Elcock, A. H. *J. Phys. Chem. Lett.* **2011**, *2*, 19–24.
- (20) (a) Wang, J.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049–1074. (b) Sorin, E. J.; Pande, V. S. *Biophys. J.* **2005**, *88*, 2472–2493.
- (21) (a) Ghosh, T.; Kalra, A.; Garde, S. *J. Phys. Chem. B* **2005**, *109*, 642–651. (b) Trzesniak, D.; Kunz, A.-P. E.; van Gunsteren, W. F. *ChemPhysChem* **2007**, *8*, 162–169. (c) Sobolewski, E.; Makowski, M.; Czaplowski, C.; Liwo, A.; Oldziej, S.; Scheraga, H. A. *J. Phys. Chem. B* **2007**, *111*, 10765–10774.
- (22) Martínez, J. M.; Pappalardo, R. R.; Marcos, E. S. *J. Am. Chem. Soc.* **1999**, *121*, 3175–3184.
- (23) Gallicchio, E.; Kubo, M. M.; Levy, R. M. *J. Phys. Chem. B* **2000**, *104*, 6271–6285.
- (24) Paschek, D. *J. Chem. Phys.* **2004**, *120*, 6674–6690.
- (25) Marcus, Y. *Biophys. Chem.* **1994**, *51*, 111.

(26) (a) Guillot, B.; Guissani, Y.; Bratos, S. *J. Chem. Phys.* **1991**, *95*, 3643–3648. (b) Guo, G.-J.; Zhang, Y.-G.; Li, M.; Wu, C.-H. *J. Chem. Phys.* **2008**, *128*, 194504.

(27) Dec, S. F.; Bowler, K. E.; Stadterman, L. L.; Koh, C. A.; Sloan, E. D. *J. Am. Chem. Soc.* **2006**, *128*, 414–415.

(28) (a) Obst, S.; Bradaczek, H. *J. Phys. Chem.* **1996**, *100*, 15677–15687. (b) Du, H.; Rasaiah, J. C.; Miller, J. D. *J. Phys. Chem. B* **2007**, *111*, 209–217.

(29) Varma, S.; Rempe, S. B. *Biophys. Chem.* **2006**, *124*, 192–199 and references therein.

(30) (a) San-Román, M. L.; Carrillo-Tripp, M.; Saint-Martin, H.; Hernández-Cobos, J.; Ortega-Blake, I. *Theor. Chem. Acc.* **2006**, *115*, 177–189. (b) Lamoureux, G.; Roux, B. *J. Phys. Chem. B* **2006**, *110*, 3308–3322. (c) Yu, H.; Whitfield, T. W.; Harder, E.; Lamoureux, G.; Vorobyov, I.; Anisimov, V. M.; Mackerell, A. D., Jr.; Roux, B. *J. Chem. Theory Comput.* **2010**, *6*, 774–786.

(31) (a) Rempe, S. B.; Pratt, L. R.; Hummer, G.; Kress, J. D.; Martin, R. L.; Redondo, A. *J. Am. Chem. Soc.* **2000**, *122*, 966–967. (b) Lyubartsev, A. P.; Laasonen, K.; Laaksonen, A. *J. Chem. Phys.* **2001**, *114*, 3120–3126.

(32) Loeffler, H. H.; Rode, B. M. *J. Chem. Phys.* **2002**, *117*, 110–117.

(33) Marcus, Y. *Chem. Rev.* **1988**, *88*, 1475–1498.

(34) (a) Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schneiders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T. *J. Phys. Chem. B* **2010**, *114*, 2549–2564. (b) Lopes, P. E. M.; Roux, B.; Mackerell, A. D., Jr. *Theor. Chem. Acc.* **2009**, *124*, 11–28.

(35) Sedláč, E.; Žoldák, G.; Antálk, M.; Sprinzl, M. *Biochim. Biophys. Acta* **2002**, *1597*, 22–27.

(36) Weisenberger, S.; Schumpe, A. *AIChE J.* **1996**, *42*, 298–300.

(37) Fennell, C. J.; Bizjak, A.; Vlachy, V.; Dill, K. A. *J. Phys. Chem. B* **2009**, *113*, 6782–6791.

(38) Otwinowski, Z.; Schevitz, R. W.; Zhang, R.-G.; Lawson, C. L.; Joachimiak, A.; Marmorstein, R. Q.; Luisi, B. F.; Sigler, P. B. *Nature* **1988**, *335*, 321–329.

Can Electron-Rich π Systems Bind Anions?

Inacrist Geronimo, N. Jiten Singh,* 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

ABSTRACT: In general, anion- π interactions exist between anions and aromatics with a positive quadrupole moment. The interaction between anions and aromatics with a negative quadrupole moment is expected to be unstable due to Coulombic repulsion. However, here we investigated the cases of aromatics with a negative quadrupole moment such as electron-rich alkyl/alkenyl/alkynyl-substituted benzenes and triphenylene, which interact with halides. Favorable binding was demonstrated with coupled cluster theory with singles, doubles, and perturbative triples excitations [CCSD(T)] at the complete basis set (CBS) limit. Stability increases with chain length, unsaturation, and halogenation. Energy decomposition analysis based on symmetry adapted perturbation theory (SAPT) shows that electrostatic repulsion is overcome by induction effects arising from the alkyl substituents.

INTRODUCTION

The design of anion receptors is traditionally predicated by interaction of the negatively charged anion with either positively charged or hydrogen bond donor groups or by coordination with metals.¹ Prime examples of this are imidazolium,² the urea complex,³ and quaternary ammonium-salt-based⁴ and calix-[*n*]arene/pyrrole receptors.⁵ Aromatic moieties are a mainstay of anion receptors as a molecular scaffold. For the anion- π interaction,⁶ the π system has generally positive traceless quadrupole moment [$\Theta_{zz} = (3Q_{zz} - \text{Tr}\{Q\})/2$], and thus it is attractive. However, the anion- π interaction when the π system has a negative quadrupole moment is expected to be repulsive because of Coulombic repulsion, so that anion recognition by electron-rich aromatic systems has hardly been explored. Experiments⁷⁻¹¹ and theoretical calculations¹²⁻¹⁵ on the aromatic systems with a positive quadrupole moment demonstrated attractive interactions for the anion. Recent experimental evidence of anion- π interaction was demonstrated in halide recognition via a copper(II)-azadendtriz complex (containing pyridine units),⁸ tetraoxacalix[2]arene[2]triazine receptors,⁹ pentafluorobenzyl-substituted ammonium and pyridinium salts,¹⁰ and Ag(I)/Cu(I)-tetrazine complexes.¹¹ Desirable properties of π interaction, including directionality and ease of fine-tuning properties through substituents, serve as an impetus in developing anion receptors based on anion- π interaction.

Well-studied model systems include the halide complexes of hexafluorobenzene ($\Theta_{zz} = 9.50 \text{ D}\text{\AA}$), 1,3,5-trinitrobenzene ($\Theta_{zz} = 22 \text{ D}\text{\AA}$), 1,3,5-tricyanobenzene ($\Theta_{zz} = 19.53 \text{ D}\text{\AA}$), and the heteroaromatic *s*-triazine ($\Theta_{zz} = 0.90 \text{ D}\text{\AA}$).¹²⁻¹⁵ In some cases, the total interaction energy of anion- π complexes is comparable to that of cation- π complexes;¹⁶ for instance, the MP2/aug-cc-pVDZ binding energy for $\text{C}_6\text{F}_6-\text{F}^-$ is 18.4 kcal/mol,^{14a} while that of $\text{C}_6\text{H}_6-\text{Na}^+$ is 22.3 kcal/mol.^{16a} Frequency calculations show that only the Cl^- , Br^- , and NO_3^- complexes of triazine and hexafluorobenzene, the Br^- complexes of 1,3,5-tricyanobenzene, and the CN^- complexes of hexafluorobenzene are actual minima in the gas phase.¹⁷ It has been demonstrated that halides preferentially form either a covalent σ or H bond complex as a consequence of electron-withdrawing groups, which increases

the acidity of aryl C-H and activates the ring toward nucleophilic substitution. This is supported by a survey of crystal structures of a neutral six-membered aromatic ring in the Cambridge Structural Database (CSD) wherein 84% of the halides are closer to the ring C than the centroid.¹⁷ A later study¹⁸ defining stringent criteria for the anion- π interaction, specifically the ring atom-anion distance of $\leq \text{vdw} + 0.2 \text{ \AA}$ and the tilt angle of $90 \pm 10^\circ$ (from the ring plane), did not yield a convincing example of anion- π interaction as the aromatic rings are invariably bound to a cationic site. A preliminary study of Cl^- complexes shows that the majority of samples have Cl^- at θ values $< 20^\circ$, which is consistent with the H bonding motif.

While most models of anion- π complexes are not actual minima, theoretical calculations are nevertheless important in shedding light on the nature of anion- π interactions. A symmetry-adapted perturbational theory (SAPT) analysis of complexes of tetrafluoroethene, hexafluorobenzene, and triazine with halides (F^- , Cl^- , Br^-) and linear (CN^-) and trigonal planar (NO_3^- , CO_3^{2-}) anions shows that the most significant contributions to the interaction energy come from electrostatic and induction effects.^{14a} As in cation- π complexes, dispersion effects are relatively low in comparison to the other energy components. Electrostatic effects dominate at large distances, but induction effects become more significant as the distance decreases. The induction contribution is attributed to the interaction of the occupied p orbital of halides or the π orbital of organic anions with the lowest unoccupied molecular orbital (LUMO) of the aromatic ring.

As stated above, anion binding with aromatics of negative Θ_{zz} is expected to be unstable due to Coulombic repulsion. However, it has already been demonstrated^{15,19} that there is no correlation between Θ_{zz} and the anion binding enthalpy, unlike the case of cation complexes, which implies that anion- π interaction is, in fact, possible for aromatics with negative Θ_{zz} if the induction overcompensates for the electrostatic repulsion. Recently, experimental evidence of anion- π interaction between electron-rich

Received: November 28, 2010

Published: February 28, 2011

alkylbenzene rings and F^- , in combination with $(C-H)^+ \cdots F^-$ -type ionic bonds, was reported for an imidazolium cage receptor.²⁰ It was particularly noted that the nonalkylated analogue cannot host F^- in the cavity between the two benzene rings that form a sandwich complex. The CCSD(T)/CBS calculations on the triethylbenzene moiety yielded an interaction energy of -0.9 kcal/mol.²⁰ A theoretical study by Wheeler and Houk¹⁵ on model $Cl-C_6H_{6-n}X_n$ ($n = 0-4, 6$; X is either electron-donating or electron-withdrawing) complexes showed that the interaction energy for trimethylbenzene is slightly attractive (the interaction energy of -0.4 kcal/mol at the M06-2X/6-31+G(d) level). More importantly, it was noted that anion binding can be attributed to direct interaction of the anion with the local dipole induced by the substituents and not with substituent-induced polarization of the π density as previously asserted.¹³ This is consistent with the results of Kim et al.,^{14a} showing a relatively low dispersion contribution in the anion binding of π systems such as hexafluorobenzene. The previous studies suggest that substituents have an important role in overcoming Coulombic repulsion between the negative Θ_{zz} of the aromatic ring and the anion. Moreover, it indicates that the nature of interaction in anion complexes of aromatics with negative Θ_{zz} is different from that with positive Θ_{zz} . Even in π - π interactions, significant differences in geometry and binding energy between electron-deficient and electron-rich aromatics have been observed.²¹

In line with this, the present study investigates the nature of anion- π interaction in aromatic systems with electron-donating alkyl substituents and an electron-withdrawing halogenated-alkyl/alkenyl/alkynyl substituent. Triply substituted benzene was chosen for the study to limit the variable factor to substituent effects. The influence of chain length, unsaturation, and halogenation in the substituent on the electrostatic, induction, dispersion, and exchange-correlation energies will be systematically studied by using high-level ab initio and SAPT calculations. The substituents include methyl (Me), ethyl (Et), $-CH=CH_2$, $-C\equiv CH$, $-CH_2F$, and $-C\equiv CF$, and a fused ring system is considered as well. The results will be discussed in comparison to anion- π complexes of trifluorobenzene, tribromobenzene, and other systems with positive Θ_{zz} . The choice of specific geometry with the anion lying above the ring center was made in order to examine and understand anion interactions with the ring electron density because of its plausible experimental realization in liquid or crystal phases. Despite the fact that the gas phase frequency calculations for such isolated systems of alkylbenzene moieties and F^- yield saddle points, it should be noted that the anion- π type interaction between alkylbenzene moieties and F^- is experimentally realized.²⁰

COMPUTATIONAL METHOD

The resolution of identity approximation of the second-order Møller-Plesset perturbation theory (RIMP2) using the aug-cc-pVDZ (aVDZ) basis set with basis set superposition error (BSSE) correction was used to optimize the geometries of various anion- π complexes. The geometry was constrained to C_{3v} symmetry so that the center of the anion lies along the C_6 axis of the aromatic ring. Vibrational frequency calculations were performed at the same level, while quadrupole moments of the RIMP2/aVDZ geometries were determined at the RHF/6-311G** level. Single point energy calculations were subsequently performed at the RIMP2/aug-cc-pVTZ (aVTZ) and CCSD(T)/

aVDZ levels with BSSE correction to obtain energies at the complete basis set (CBS) limit. The MP2 CBS limit was evaluated by using the extrapolation scheme based on the proportionality of the basis set error in the electron correlation energy to N^{-3} for the aug-cc-pVNZ basis set.²² This was then used to estimate the CCSD(T)/CBS limit by adding the CCSD(T)/aVDZ binding energies and the difference between the MP2/CBS and MP2/aVDZ binding energies.

The total interaction energy was decomposed into electrostatic (E_{es}), induction (E_{ind}), dispersion (E_{disp}), and exchange-repulsion (E_{exch}) components, or effective induction (E_{in}), effective dispersion (E_{dp}), and effective exchange-repulsion (E_x) components based on SAPT,²³ as described previously.²⁴

$$E_{tot} = E_{es} + E_{in} + E_{dp} + E_x$$

where

$$E_{es} = E_{es}^{(10)} + E_{es,resp}^{(12)}$$

$$E_{in} = E_{ind}^{(20)} + E_{exch,ind,resp}^{(20)} + \delta_{ind,resp}^{HF} + {}^t E_{ind}^{(22)} + {}^t E_{exch,in}^{(22)} - 0.4 \\ \times (E_{CCSD(T)/CBS} - E_{SAPT(MP2)/aVDZ'})$$

$$E_{dp} = E_{disp}^{(20)} + E_{exch,disp}^{(20)} + \delta_{ind,resp}^{HF} - 0.6$$

$$\times (E_{CCSD(T)/CBS} - E_{SAPT(MP2)/aVDZ'})$$

$$E_x = E_{exch}^{(10)} + E_{exch}^{(11)} + E_{exch}^{(12)}$$

The superscripts refer to orders in the intermolecular interactions and intramolecular correlation potential, “resp” to the inclusion of coupled Hartree-Fock response, and $\delta_{ind,resp}^{HF}$ to higher-order Hartree-Fock induction and exchange-induction contributions. E_{in} is the sum of the exchange-induction term and E_{ind} . E_{dp} is the sum of the exchange-dispersion term and E_{disp} , while E_x excludes the aforementioned terms from E_{exch} . E_{dp} and E_{in} include the difference between the CCSD(T)/CBS and SAPT(MP2)/aVDZ' binding energies to correct for the basis set dependency of the dispersion and induction energy, as discussed in previous studies.²⁴ SAPT calculations were performed with SAPT2008²⁵ at the MP2/aVDZ' level where the p diffuse functions on H and the d diffuse functions on heavy atoms are removed. RIMP2 and CCSD(T) calculations were done using Turbomole 6.0.2²⁶ and Gaussian 09,²⁷ respectively.

RESULTS AND DISCUSSION

Optimized geometries of the F^- complexes of various substituted benzenes, constrained to C_{3v} symmetry, are shown in Figure 1. The complete list of traceless quadrupole moments (Θ_{zz}), perpendicular distance (R_c), and interaction energies are summarized in Table 1, while SAPT components of the interaction energy of selected complexes are listed in Table 2. The focus of the present work is to determine the effect of electron-donating alkyl substituents on the anion- π type interaction; hence these geometries will be used in the discussion despite the fact that these structures are saddle points.

Bz-F, with an interaction energy of 0.66 kcal/mol, is unstable with respect to the dissociated F^- and benzene. However, the F^- - π complex becomes stable with methyl (-0.38 kcal/mol)

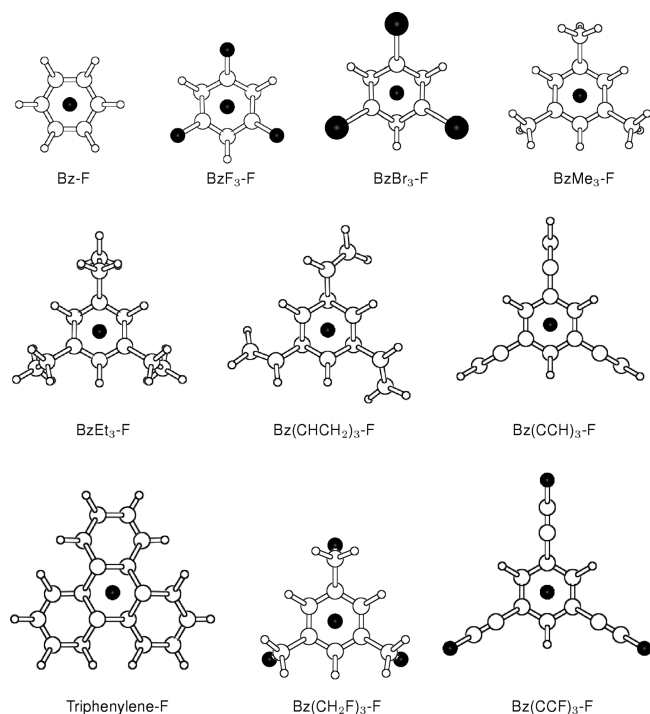


Figure 1. Optimized geometries of various F^- - π complexes obtained at the RIMP2/aVDZ level with BSSE correction.

and ethyl (-0.86 kcal/mol) substitution. The increased stability from $Bz-F$ to $BzEt_3-F$ is accompanied by a decrease in R_v , which caused increased exchange repulsion with chain length. While Θ_{zz} is the least negative in $BzMe_3$, the repulsive electrostatic contribution increases uniformly with chain length, which indicates that the charge-quadrupole interaction is not significant. Wheeler and Houk¹⁵ attributed this to the short distance between the halide and aromatic ring, in which case the expansion of intermolecular electrostatic interactions in terms of electric multipoles cannot be considered valid or accurate. The attractive inductive effects increase from $Bz-F$ to $BzEt_3-F$ due to anion-induced polarization of alkyl substituents, which becomes significant with increasing chain length. Clements and Lewis¹⁹ also showed that favorable anion binding of halo-substituted aromatics arises from the polarizability of the substituents as evidenced by the increase in anion binding with increased number and polarizability of the substituent. The attractive dispersion contribution also increases but is lower than those of other components.

The effect of unsaturation is determined by a comparison of $BzEt_3-F$, $Bz(CHCH_2)_3-F$, and $Bz(CCH)_3-F$. The F^- and Cl^- complexes of $Bz(CCH)_3$ were reported earlier by Wheeler and Houk¹⁵ and Lucas et al.^{13e,f} The increase in stability is dramatic, from -0.86 kcal/mol in $BzEt_3-F$ to -7.11 kcal/mol in $Bz(CCH)_3-F$. The system is further stabilized in fused rings (triphenylene- F), as can be seen from the RIMP2/CBS energies (-7.34 kcal/mol compared to -6.58 kcal/mol in $Bz(CCH)_3-F$). Θ_{zz} becomes more negative with unsaturation, and yet, there is a shift from repulsive to attractive electrostatic energy from $BzEt_3-F$ to $Bz(CCH)_3-F$. The attractive electrostatic term in $Bz(CCH)_3-F$ indicates that the ethynyl group is electron-withdrawing. The increase in attractive induction effects is significant upon unsaturation, while that for dispersion is moderate.

Table 1. BSSE-Corrected RIMP2 and CCSD(T) Interaction Energies (in kcal/mol) on the BSSE-Corrected RIMP2/aVDZ Optimized Geometries^a

complex	Θ_{zz}	R_v	MP2			CCSD(T)	
			aVDZ	aVTZ	CBS	aVDZ	CBS
$Bz-F$	-8.76	3.270	1.72	1.08	0.81	1.57	0.66
$Bz-Cl$	-8.76	3.812	1.01	0.44	0.20	1.37	0.56
$Bz-Br$	-8.76	3.870	0.86	0.20	-0.08	1.38	0.44
$BzMe_3-F$	-8.46	3.157	0.88	0.17	-0.14	0.63	-0.38
$BzEt_3-F$	-9.11	3.080	0.24	-0.28	-0.50	-0.11	-0.86
$BzEt_3-Cl$	-9.11	3.569	-0.54	-1.11	-1.35	-0.05	-0.86
$BzEt_3-Br$	-9.11	3.730	-0.73	-1.34	-1.59	-0.11	-0.97
$Bz(CHCH_2)_3-F$	-15.37	2.829	-1.95	-3.19	-3.71	-2.00	-3.76
$Bz(CCH)_3-F$	-16.99	2.612	-5.47	-6.25	-6.58	-6.00	-7.11
$Bz(CCH)_3-Cl$	-16.99	3.293	-3.71	-4.59	-4.97	-2.95	-4.20
$Bz(CCH)_3-Br$	-16.99	3.486	-3.30	-4.20	-4.58	-2.35	-3.64
triphenylene- F	-25.16	2.439	-6.01	-6.95	-7.34	NA	NA
triphenylene- Cl	-25.16	3.117	-3.26	-4.43	-4.92	NA	NA
BzF_3-F	0.69	2.755	-8.20	-8.68	-8.88	-8.62	-9.31
BzF_3-Cl	0.69	3.330	-6.26	-6.91	-7.19	-5.63	-6.56
BzF_3-Br	0.69	3.517	-5.72	-6.40	-6.69	-4.98	-5.95
$BzBr_3-F$	-3.91	2.658	-10.34	-11.56	-12.07	-10.58	-12.31
$Bz(CH_2F)_3-F$	-6.43	2.736	-13.78	-14.28	-14.48	-13.85	-14.55
$Bz(CH_2F)_3-Cl$	-6.43	3.346	-12.72	-13.35	-13.62	-11.83	-12.74
$Bz(CH_2F)_3-Br$	-6.43	3.516	-12.32	-13.00	-13.28	-11.31	-12.27
$Bz(CCF)_3-F$	-2.65	2.709	-7.56	-8.00	-8.18	-8.67	-9.29
$Bz(CCF)_3-Cl$	-2.65	3.290	-6.00	-6.51	-6.73	-5.45	-6.18
$Bz(CCF)_3-Br$	-2.65	3.479	-5.57	-6.10	-6.33	-4.82	-5.58

^a Θ_{zz} : quadrupole moment of the uncomplexed aromatic system (in $D\text{\AA}$) calculated at the RHF/6-311G** level, R_v : perpendicular distance between the ring centroid and anion in \AA . Geometries are constrained to C_{3v} symmetry. Vibrational frequency calculations demonstrate that only BzF_3-Br is a genuine minimum, while $BzMe_3-F$ has 5 imaginary frequencies and the rest have 2 each. Positive interaction energy indicates that the complex is less stable than the dissociated structure.

Table 2. SAPT(MP2)/aVDZ' Energies (in kcal/mol) Calculated Using the BSSE-Corrected RIMP2/aVDZ Geometries

complex	E_{tot}	E_{es}	E_{in}	E_{dp}	E_x
$Bz-F$	0.66	5.69	-7.13	-1.58	3.67
$Bz-Cl$	0.56	3.86	-4.08	-2.29	3.07
$Bz-Br$	0.44	2.99	-3.78	-2.94	4.17
$BzMe_3-F$	-0.38	5.92	-9.41	-1.88	4.99
$BzEt_3-F$	-0.86	6.24	-11.07	-2.11	6.08
$BzEt_3-Cl$	-0.86	3.71	-6.61	-3.72	5.76
$BzEt_3-Br$	-0.97	2.89	-5.65	-4.25	6.04
$Bz(CHCH_2)_3-F$	-3.76	2.07	-14.77	-2.62	11.56
$Bz(CCH)_3-F$	-7.11	-5.66	-17.82	-3.70	20.08
BzF_3-F	-9.31	-10.03	-11.46	-1.95	14.14
BzF_3-Cl	-6.56	-7.53	-5.54	-3.75	10.26
BzF_3-Br	-5.95	-7.19	-4.50	-4.17	9.90
$BzBr_3-F$	-12.31	-10.59	-16.44	-3.34	18.07
$Bz(CH_2F)_3-F$	-14.55	-15.34	-13.10	-1.38	15.28
$Bz(CH_2F)_3-Cl$	-12.74	-13.46	-6.15	-3.53	10.41
$Bz(CCF)_3-F$	-9.29	-5.98	-15.88	-3.12	15.69

Halogenation of substituents increases stability, as demonstrated by $Bz-F/BzF_3-F$ ($0.66/-9.31$ kcal/mol), $BzMe_3-F/Bz(CH_2F)_3-F$ ($-0.38/-14.55$ kcal/mol), and $Bz(CCH)_3-F/Bz(CCF)_3-F$ ($-7.11/-9.29$ kcal/mol). While the Θ_{zz} values of $Bz(CH_2F)_3$ and $Bz(CCF)_3$ are still negative, they are

more positive than their unhalogenated counterparts. There is a shift from repulsive to attractive electrostatic terms and a significant increase in exchange-repulsion due to decreased R_v in the case of $Bz-F/BzF_3-F$ and $BzMe_3-F/Bz(CH_2F)_3-F$. The attractive induction term also increases, similar to the effect of chain length and unsaturation. Attractive dispersion effects increase from $Bz-F$ to BzF_3-F but decrease from $Bz(Me)_3-F$ to $Bz(CH_2F)_3-F$. On the other hand, the attractive electrostatic term from $Bz(CCH)_3-F$ to $Bz(CCF)_3-F$ increases only slightly, while attractive induction and dispersion effects, as well as exchange-repulsion, decrease.

Electrostatic and induction effects are comparable in magnitude in alkyl-substituted benzenes (Bz , $BzMe_3$, $BzEt_3$, and $Bz(CH_2F)_3$) like the case of anion- π complexes of electron-deficient aromatic rings such as hexafluorobenzene, triazine,^{14a} BzF_3 , and $BzBr_3$. However, the electrostatic term is repulsive for Bz , $BzMe_3$, and $BzEt_3$. In the case of unsaturated substituents as in $Bz(CHCH_2)_3$, $Bz(CCH)_3$, and $Bz(CCF)_3$, the attractive induction term is much larger than the electrostatic term. Exchange-repulsion terms have the highest magnitude in the case of hexafluorobenzene and triazine but are comparable to the electrostatic and/or inductive effects in alkyl-substituted benzenes, BzF_3 , and $BzBr_3$. In both electron-deficient and alkyl-substituted benzenes, dispersion effects are relatively low. Unlike the halide complexes of hexafluorobenzene, triazine,^{14a} BzF_3 , $Bz(CCH)_3$, $Bz(CH_2F)_3$, and $Bz(CCF)_3$, the interaction energies for the halide complexes of Bz and $BzEt_3$ become more attractive from F^- to Br^- . The electrostatic energy becomes more attractive in the latter, while the opposite occurs in hexafluorobenzene, triazine,^{14a} BzF_3 , and $Bz(CH_2F)_3$. The induction energy becomes more repulsive in Bz , $BzEt_3$, BzF_3 , and $Bz(CH_2F)_3$ and attractive in hexafluorobenzene.^{14a} This parallels the decrease in electronegativity and the increase in polarizability of the anion from F^- to Br^- . The attractive dispersion energy increases in all cases.

A Cambridge Structural Database (CSD, version 5.31, November 2009) search on possible anion- π interactions between halides, NO_3^- , or ClO_4^- and alkylated arenes with an anion-centroid distance R_v of 5 Å and a tilt angle θ of $90 \pm 15^\circ$ yielded 32, 1, and 4 samples, respectively. All fragments are bonded to positively charged groups. Four samples, all interacting with Br^- , are found to contain trialkylated benzene fragments — RAJXUW ($R=Me$),²⁸ REQYOC ($R=Ph$),²⁹ and UMICEY/UMICOI ($R=Et$).³⁰ Only two samples, BIHQOZ,³¹ a macrobicyclic azaphane receptor, and REQYOC,³⁵ an N-spiro chiral quaternary ammonium bromide, satisfy the following criteria based on optimized geometries of halide complexes of $BzEt_3$ ($R_v = 3-4$ Å, ring atom-anion distance of $\leq vdw + 0.5$ Å and $\theta = 90^\circ \pm 10^\circ$). In the former, anion recognition is primarily based on CH and NH anion interactions. However, the I^- -centroid distance is 3.692 Å, and all tilt angles are 90° , which is consistent with theoretical calculations on anion- π complexes.

CONCLUSION

The present study demonstrates the presence of anion- π interactions between halides and alkyl/alkenyl/alkynyl-substituted aromatics. An increase in the chain length, unsaturation, and halogenation of the alkyl substituents is paralleled by an increase in stability of the anion- π complex. Despite the repulsive electrostatic term, the F^- complexes of $BzMe_3$, $BzEt_3$, $Bz(CHCH_2)_3$, and $Bz(CCH)_3$ are stable due to attractive

induction effects arising from the substituent. The attractive dispersion contributions are relatively low. The F^- complexes of BzF_3 , $BzBr_3$, $Bz(CH_2F)_3$, and $Bz(CCF)_3$ have an attractive electrostatic term, but these are lower in magnitude than the induction energy, except in the case of $Bz(CH_2F)_3$. Unlike F^- complexes of aromatics with a positive Θ_{zz} like hexafluorobenzene and triazine, the exchange-repulsion is comparable in magnitude to these two terms. The interaction energies become more repulsive from F^- to Br^- in BzF_3 , $Bz(CCH)_3$, $Bz(CH_2F)_3$, and $Bz(CCF)_3$ as in the case of hexafluorobenzene and triazine^{14a} and more attractive in Bz and $BzEt_3$. The attractive anion- π interactions in alkyl-substituted aromatic complexes can be utilized in the design of anion receptors as already demonstrated by imidazolium-based receptors.²⁰

AUTHOR INFORMATION

Corresponding Author

*E-mail: kim@postech.ac.kr (K.S.K.), jiten@postech.ac.kr (N.J.S.).

ACKNOWLEDGMENT

This work was supported by NRF (WCU: R32-2008-000-10180-0; National Honor Scientist Program: 2010-0020414) and KISTI (KSC-2008-K08-0002).

REFERENCES

- (1) (a) Caltagirone, C.; Gale, P. A. *Chem. Soc. Rev.* **2009**, 38, 520–563. (b) Singh, N. J.; Olleta, A. C.; Kumar, A.; Park, M.; Yi, H.-B.; Bandyopadhyay, I.; Lee, H. M.; Tarakeshwar, P.; Kim, K. S. *Theor. Chem. Acc.* **2006**, 115, 127–135. (c) Singh, N. J.; Lee, H. M.; Hwang, I.-C.; Kim, K. S. *Supramol. Chem.* **2007**, 19, 321–332.
- (2) (a) Yoon, J.; Kim, S. K.; Singh, N. J.; Kim, K. S. *Chem. Soc. Rev.* **2006**, 35, 355–360. (b) Chellapan, K.; Singh, N. J.; Hwang, I. C.; Lee, J. W.; Kim, K. S. *Angew. Chem., Int. Ed.* **2005**, 44, 2899–2903. (c) Yun, S.; Ihm, H.; Kim, H. G.; Lee, C.-W.; Indrajit, B.; Oh, K. S.; Gong, Y. J.; Lee, J. W.; Yoon, J.; Lee, H. C.; Kim, K. S. *J. Org. Chem.* **2003**, 68, 2467–2470. (d) Ihm, H.; Yun, S.; Kim, H. G.; Kim, J. K.; Kim, K. S. *Org. Lett.* **2002**, 4, 2897–2900.
- (3) (a) dos Santos, C. M. G.; Fernández, P. B.; Plush, S. E.; Leonard, J. P.; Gunnlaugsson, T. *Chem. Commun.* **2007**, 3389–3391. (b) Kim, S. K.; Singh, N. J.; Kim, S. J.; Swamy, K. M. K.; Kim, S. H.; Lee, K.-H.; Kim, K. S.; Yoon, J. *Tetrahedron* **2005**, 61, 4545–4550.
- (4) Luxami, V.; Sharma, N.; Kumar, S. *Tetrahedron Lett.* **2008**, 49, 4265–4268.
- (5) (a) Prados, P.; Quesada, R. *Supramol. Chem.* **2008**, 20, 201–216. (b) Hong, B. H.; Lee, J. Y.; Lee, C.-W.; Kim, J. C.; Bae, S. C.; Kim, K. S. *J. Am. Chem. Soc.* **2001**, 123, 10748–10749.
- (6) (a) Schneider, H.-J.; Schiestel, T.; Zimmermann, P. *J. Am. Chem. Soc.* **1992**, 114, 7698–7703. (b) Schneider, H.-J.; Schiestel, T.; Zimmermann, P. *J. Phys. Org. Chem.* **1993**, 6, 590–594. (c) Schottel, B. L.; Chifotides, H. T.; Dunbar, K. R. *Chem. Soc. Rev.* **2008**, 37, 68–83.
- (7) (a) Rosokha, Y. S.; Lindeman, S. V.; Rosokha, S. V.; Kochi, J. K. *Angew. Chem., Int. Ed.* **2004**, 43, 4650–4652. (b) de Hoog, P.; Gamez, P.; Mutikainen, I.; Turpeinen, U.; Reedijk, J. *Angew. Chem., Int. Ed.* **2004**, 43, 5815–5817. (c) Demeshko, S.; Dechert, S.; Meyer, F. *J. Am. Chem. Soc.* **2004**, 126, 4508–4509. (d) Campos-Fernández, C. S.; Schottel, B. L.; Chifotides, H. T.; Bera, J. K.; Bacsá, J.; Koomen, J. M.; Russell, D. H.; Dunbar, K. R. *J. Am. Chem. Soc.* **2005**, 127, 12909–12923. (e) Mascal, M.; Yakovlev, I.; Nikitin, E. B.; Fettingier, J. C. *Angew. Chem., Int. Ed.* **2007**, 46, 8782–8784. (f) Hay, B. P.; Bryantsev, V. S. *Chem. Commun.* **2008**, 2417–2428. (g) 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, 533–538.

- (8) de Hoog, P.; Robertazzi, A.; Mutikainen, I.; Turpeinen, U.; Gamez, P.; Reedijk, J. *Eur. J. Inorg. Chem.* **2009**, 2684–2690.
- (9) Wang, D.-X.; Zheng, Q.-Y.; Wang, Q.-Q.; Wang, M.-X. *Angew. Chem., Int. Ed.* **2008**, *47*, 7485–7488.
- (10) Albrecht, M.; Müller, M.; Mergel, O.; Rissanen, K.; Valkonen, A. *Chem.—Eur. J.* **2010**, *16*, 5062–5069.
- (11) Gural'skiy, I. A.; Escudero, D.; Frontera, A.; Solntsev, P. V.; Rusanov, E. B.; Chernega, A. N.; Krautscheidd, H.; Domasevitch, K. V. *Dalton Trans.* **2009**, 2856–2864.
- (12) (a) Mascal, M.; Armstrong, A.; Bartberger, M. D. *J. Am. Chem. Soc.* **2002**, *124*, 6274–6276. (b) Alkorta, I.; Rozas, I.; Elguero, J. *J. Am. Chem. Soc.* **2002**, *124*, 8593–8598.
- (13) (a) Quiñonero, D.; Garau, C.; Rotger, C.; Frontera, A.; Ballester, P.; Costa, A.; Deyà, P. M. *Angew. Chem., Int. Ed.* **2002**, *41*, 3389–3392. (b) Garau, C.; Frontera, A.; Quiñonero, D.; Ballester, P.; Costa, A.; Deyà, P. M. *J. Phys. Chem. A* **2004**, *108*, 9423–9427. (c) Quiñonero, D.; Garau, C.; Frontera, A.; Ballester, P.; Costa, A.; Deyà, P. M. *J. Phys. Chem. A* **2005**, *109*, 4632–4637. (d) Frontera, A.; Quiñonero, D.; Costa, A.; Ballester, P.; Deyà, P. M. *New J. Chem.* **2007**, *31*, 556–560. (e) Lucas, X.; Quiñonero, D.; Frontera, A.; Deyà, P. M. *J. Phys. Chem. A* **2009**, *113*, 10367–10375. (f) Lucas, X.; Frontera, A.; Quiñonero, D.; Deyà, P. M. *J. Phys. Chem. A* **2010**, *114*, 1926–1930.
- (14) (a) Kim, D.; Tarakeshwar, P.; Kim, K. S. *J. Phys. Chem. A* **2004**, *108*, 1250–1258. (b) Kim, D. Y.; Singh, N. J.; Lee, J. W.; Kim, K. S. *J. Chem. Theory Comput.* **2008**, *4*, 1162–1169. (c) Kim, D. Y.; Singh, N. J.; Kim, K. S. *J. Chem. Theory Comput.* **2008**, *4*, 1401–1407.
- (15) Wheeler, S. E.; Houk, K. N. *J. Phys. Chem. A* **2010**, *114*, 8658–8664.
- (16) (a) Singh, N. J.; Min, S. K.; Kim, D. Y.; Kim, K. S. *J. Chem. Theory Comput.* **2009**, *5*, 515–529. (b) Kim, D.; Hu, S.; Tarakeshwar, P.; Kim, K. S.; Lisy, J. M. *J. Phys. Chem. A* **2003**, *107*, 1228–1238. (c) Choi, H. S.; Suh, S. B.; Cho, S. J.; Kim, K. S. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 12094–12099. (d) Dougherty, D. A.; Stauffer, D. *Science* **1990**, *250*, 1558–1560.
- (17) Berryman, O. B.; Bryantsev, V. S.; Stay, D. P.; Johnson, D. W.; Hay, B. P. *J. Am. Chem. Soc.* **2007**, *129*, 48–58.
- (18) Hay, B. P.; Custelcean, R. *Cryst. Growth Des.* **2009**, *9*, 2539–2545.
- (19) Clements, A.; Lewis, M. *J. Phys. Chem. A* **2006**, *110*, 12705–12710.
- (20) Xu, Z.; Singh, N. J.; Kim, S. K.; Spring, D. R.; Kim, K. S.; Yoon, J. *Chem.—Eur. J.* **2010**, *17*, 1163–1170.
- (21) (a) Riley, K. E.; Pitonak, M.; Jurecka, P.; Hobza, P. *Chem. Rev.* **2010**, *110*, 5023–5063. (b) Lee, E. C.; Kim, D.; Jurecka, P.; Tarakeshwar, P.; Hobza, P.; Kim, K. S. *J. Phys. Chem. A* **2007**, *111*, 3446–3457. (c) Kim, K. S.; Tarakeshwar, P.; Lee, J. Y. *Chem. Rev.* **2000**, *100*, 4145–4185.
- (22) (a) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *J. Chem. Phys.* **1997**, *106*, 9639–9646. (b) Min, S. K.; Lee, E. C.; Lee, H. M.; Kim, D. Y.; Kim, D.; Kim, K. S. *J. Comput. Chem.* **2008**, *29*, 1208–1221.
- (23) Jeziorski, B.; Moszynski, R.; Szalewicz, K. *Chem. Rev.* **1994**, *94*, 1887–1930.
- (24) Singh, N. J.; Min, S. K.; Kim, D. Y.; Kim, K. S. *J. Chem. Theory Comput.* **2009**, *5*, 515–529.
- (25) Bukowski, R.; Cencek, W.; Jankowski, P.; Jeziorska, M.; Jeziorski, B.; Kucharski, S. A.; Lotrich, V. F.; Misquitta, A. J.; Moszynski, R.; Patkowski, K.; Podeszwa, R.; Rybak, S.; Szalewicz, K.; Williams, H. L.; Wheatley, R. J.; Wormer, P. E. S.; Zuchowski, P. S. *SAPT2008*; University of Delaware: Newark, DE, 2008. See also ref 23.
- (26) *TURBOMOLE V6.0 2009*; University of Karlsruhe; Forschungszentrum Karlsruhe GmbH: Karlsruhe, Germany, 2007. Available from <http://www.turbomole.com> (accessed Feb 2011).
- (27) 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.02; Gaussian, Inc.: Wallingford, CT, 2009.
- (28) Baker, M. V.; Bosnich, M. J.; Brown, D. H.; Byrne, L. T.; Hesler, V. J.; Skelton, B. W.; White, A. H.; Williams, C. C. *J. Org. Chem.* **2004**, *69*, 7640–7652.
- (29) Ooi, T.; Uematsu, Y.; Kameda, M.; Maruoka, K. *Tetrahedron* **2006**, *62*, 11425–11436.
- (30) Wallace, K. J.; Belcher, W. J.; Turner, D. R.; Syed, K. F.; Steed, J. W. *J. Am. Chem. Soc.* **2003**, *125*, 9699–9715.
- (31) Ilioudis, C. A.; Tocher, D. A.; Steed, J. W. *J. Am. Chem. Soc.* **2004**, *126*, 12395–12402.

Resolutions of the Coulomb Operator: IV. The Spherical Bessel Quasi-Resolution

Taweetham Limpanuparb,* Andrew T. B. Gilbert, and Peter M. W. Gill

Research School of Chemistry, Australian National University, Canberra ACT 0200, Australia

ABSTRACT: We show that the Coulomb operator can be resolved as $r_{12}^{-1} = \sum_{nlm} \phi_{nlm}(r_1) \phi_{nlm}(r_2)$ where $\phi_{nlm}(r)$ is proportional to the product of a spherical Bessel function and a spherical harmonic, provided that $r_1 + r_2 < 2\pi$. The resolution reduces Coulomb matrix elements to Cholesky-like sums of products of auxiliary integrals. We find that these sums converge rapidly for four prototypical electron densities. To demonstrate its viability in large-scale quantum chemical calculations, we also use a truncated resolution to calculate the Coulomb energy of the nanodiamond crystallite $C_{84}H_{64}$.

The apparently innocuous Coulomb operator

$$r_{12}^{-1} \equiv |\mathbf{r}_1 - \mathbf{r}_2|^{-1} \quad (1)$$

lies at the heart of many of challenging problems in contemporary quantum chemistry, and many ingenious schemes have been devised^{1–17} to treat it efficiently and accurately. In most cases, the full complexity of the operator is avoided by partially decoupling it^{1–3} and employing multipole expansions,^{4–7} Fourier transforms,⁸ Cholesky decomposition,^{9–12} density fitting,^{13–15} or other such methods.^{16,17}

Our contributions^{18–20} employ Coulomb resolutions

$$r_{12}^{-1} = \sum_{n=0}^{\infty} \sum_{l=0}^{\infty} \sum_{m=-l}^l \phi_{nlm}(r_1) \phi_{nlm}(r_2) \quad (2)$$

where the one-particle functions

$$\phi_{nlm}(\mathbf{r}) = V_{nl}(r) Y_{lm}(\mathbf{r}) \quad (3)$$

involve a radial function $V_{nl}(r)$ and a real spherical harmonic Y_{lm} .²¹ Such resolutions reduce Coulomb matrix elements to sums of auxiliary integrals

$$\langle a | r_{12}^{-1} | b \rangle = \sum_{nlm} \langle a | \phi_{nlm} \rangle \langle \phi_{nlm} | b \rangle \quad (4)$$

and thus formally resemble Cholesky schemes.^{9–12} However, our approach forms the “Cholesky triangle” directly, without computing the matrix elements.

To construct a Coulomb resolution, one combines the Legendre expansion and the Addition Theorem²¹ to obtain the well-known²¹ angular resolution

$$r_{12}^{-1} = \sum_{lm} \frac{4\pi}{2l+1} \frac{r_{<}^l}{r_{>}^{l+1}} Y_{lm}(r_1) Y_{lm}(r_2) \quad (5)$$

where $r_{<}$ and $r_{>}$ are the smaller and larger of r_1 and r_2 .

To achieve a radial resolution

$$\frac{4\pi}{2l+1} \frac{r_{<}^l}{r_{>}^{l+1}} = \sum_n V_{nl}(r_1) V_{nl}(r_2) \quad (6)$$

one possibility^{18–20} is to choose

$$V_{nl}(r) = 2\sqrt{2} \int_0^{\infty} h_n(x) j_l(xr) dx \quad (7)$$

where the j_l are spherical Bessel functions²¹ and the h_n are any functions that form a complete and orthonormal set on $[0, \infty)$. We chose the Hermite functions

$$h_n(x) = \frac{(2/\pi)^{1/4}}{2^n \sqrt{(2n)!}} H_{2n}(x/\sqrt{2}) \exp(-x^2/4) \quad (8)$$

in our first work¹⁸ but adopted the Laguerre functions

$$h_n(x) = \sqrt{2} L_n(2x) \exp(-x) \quad (9)$$

in later studies.^{19,20} This approach to the radial resolution is theoretically attractive, but unfortunately, the radial functions V_{nl} that emerge from such “natural” choices for the h_n are often computationally expensive.^{18–20} This has led us to explore alternative schemes.

In the present letter, we offer a route based on the recently proven identity²²

$$\int_0^{\infty} j_l(nx) j_l(ny) dn \stackrel{\circ}{=} \frac{\delta_{l,0}}{2} + \sum_{n=1}^{\infty} j_l(nx) j_l(ny) \quad (10)$$

where $l = 0, 1, 2, \dots$ and $|x| + |y| < 2\pi$. We use the symbol $\stackrel{\circ}{=}$ to remind us of this domain restriction.

If we begin with the integral representation²³ of the left-hand side of eq 6

$$\frac{4\pi}{2l+1} \frac{r_{<}^l}{r_{>}^{l+1}} = 8 \int_0^{\infty} j_l(xr_1) j_l(xr_2) dx \quad (11)$$

and apply eq 10, we obtain the radial quasi-resolution

$$\frac{4\pi}{2l+1} \frac{r_{<}^l}{r_{>}^{l+1}} \stackrel{\circ}{=} 8 \left[\frac{\delta_{l,0}}{2} + \sum_{n=1}^{\infty} j_l(nr_1) j_l(nr_2) \right] \quad (12)$$

Published: March 16, 2011

Table 1. Coulomb Energies E , Components $\Delta\tilde{E}^{(n)}$, and Domain-Violation Errors^a E_{DVE} of Four Radial Charge Densities^b $\rho(r)$

	uniform density ^c	exponential density	rational density	Gaussian density
$R^3 \times \rho(r)$	$3/(4\pi) H(R-r)$	$\exp(-r/R)/(8\pi)$	$(1+(r/R)^2)^{-2}/\pi^2$	$\exp(-r^2/R^2)/\pi^{3/2}$
nonanalyticity	discontinuity at $r=R$	cusp at $r=0$	poles at $r=\pm iR$	no singularities
$R \times E$	3/5	5/32	$1/(2\pi)$	$1/(2\pi)^{1/2}$
$\Delta\tilde{E}^{(n)}$	$9j_1^2(nR)/(nR)^2$	$(1+n^2R^2)^{-4}$	$\exp(-2nR)$	$\exp(-n^2R^2/2)$
convergence	$O[(nR)^{-4}]$	$O[(nR)^{-8}]$	$O[\exp(-2nR)]$	$O[\exp(-n^2R^2/2)]$
$R \times E_{\text{DVE}}$	$6(1-\theta)^3 H(1-\theta)/\theta^4$	$1/6(\theta+1)^3 \exp(-2\theta)$	$1/(6\theta)$	$(2/\pi)^{1/2} \exp(-2\theta^2)$

^a $\theta = \pi/R$. ^b R is a parameter that characterizes the radial extent of the density. ^c H is the Heaviside step function.²¹

and then the spherical Bessel quasi-resolution

$$r_{12}^{-1} = \sum_{nlm} \phi_{nlm}(r_1) \phi_{nlm}(r_2) \quad (13)$$

where the one-particle functions are

$$\phi_{nlm}(r) = 2\sqrt{2 - \delta_{n,0}} j_l(nr) Y_{lm}(r) \quad (14)$$

This is the key result of our letter. As the prefix “quasi” and the symbol = emphasize, it is valid only for $r_1 + r_2 < 2\pi$. The quasi-resolution, unlike our previous resolutions,^{18–20} requires only the calculation of spherical Bessel functions²⁴ and spherical harmonics,²⁵ which is efficient and stable even for large n , l , and m .

Replacing r_{12}^{-1} with the quasi-resolution directly yields the Cholesky-like decomposition

$$\langle a|r_{12}^{-1}|b \rangle = \sum_{nlm} \langle a|\phi_{nlm} \rangle \langle \phi_{nlm}|b \rangle \quad (15)$$

but without the need to compute the $\langle a|r_{12}^{-1}|b \rangle$ integrals. The auxiliary integrals

$$\langle a|\phi_{nlm} \rangle = 2\sqrt{2 - \delta_{n,0}} \int a(r) j_l(nr) Y_{lm}(r) dr \quad (16)$$

are easily found if the Fourier transform of $a(r)$ is known. For example, if $a(r)$ is the Gaussian

$$a(r) = (\alpha/\pi)^{3/2} \exp(-\alpha|r-R|^2) \quad (17)$$

we have

$$\langle a|\phi_{nlm} \rangle = \exp\left(-\frac{n^2}{4\alpha}\right) \phi_{nlm}(R) \quad (18)$$

If $a(r)$ is sufficiently smooth, then, by Darboux’s principle,²⁶ the $\langle a|\phi_{nlm} \rangle$ will decay quickly for large n , l , and m , leading to rapid convergence of the sum in eq 15. We see from eq 18, for example, that small α yield fast decay with n , and small R yield fast decay with l .

One elementary use of the quasi-resolution is to find the Coulomb self-interaction energy:

$$E = \frac{1}{2} \langle \rho|r_{12}^{-1}|\rho \rangle \quad (19)$$

of a given charge density $\rho(r)$. If the density $\rho(r) \equiv \rho(r)$ is a normalized, origin-centered radial function, one finds

$$\tilde{E} = \frac{1}{2} \sum_{nlm} \langle \rho|\phi_{nlm} \rangle \langle \phi_{nlm}|\rho \rangle = \frac{1}{2\pi} + \frac{1}{\pi} \sum_n \Delta\tilde{E}^{(n)} \quad (20)$$

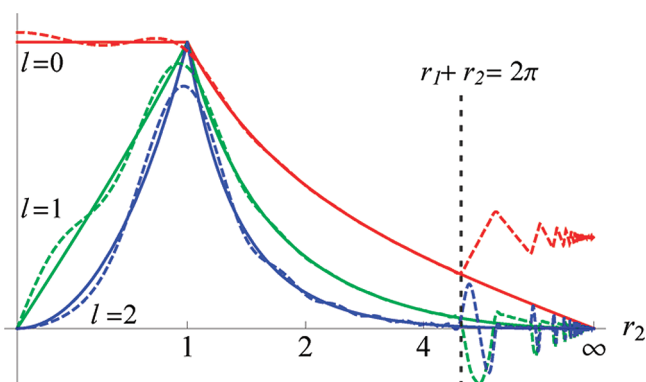


Figure 1. The left-hand side (solid) and right-hand side (dashed) of eq 12 for $l=0, 1$, and 2 when $r_1=1$ and the sum is truncated after $n=10$. Plots are scaled so that the left-hand sides coincide at $r_2=1$.

and results for four such densities are given in Table 1. These densities consist of a uniform ball (which is discontinuous on its boundary), an exponential (which has a cusp at a point), a rational function (which has poles in the complex plane), and a Gaussian (which is entire). Consistent with Darboux’s principle,²⁶ the results in the penultimate row of Table 1 confirm that the convergence of the resolution eq 20 is algebraic if $\rho(r)$ has a singularity in real space, exponential if it has a singularity in the complex plane, and superexponential if $\rho(r)$ is entire.

The key weakness of the quasi-resolution is the domain restriction $r_1 + r_2 < 2\pi$. If the quasi-resolution is applied to a density that extends beyond $r = \pi$, it introduces a domain-violation error

$$E_{\text{DVE}} = \tilde{E} - E \quad (21)$$

and the final row of Table 1 illustrates this. The message is clear: in practical applications, one should scale the system so that the DVE is acceptably small.

We begin our numerical assessment by truncating the radial resolution eq 12 after N terms. The truncated sums are useful approximations to the left-hand side, and Figure 1 illustrates this for $l=0, 1$, and 2 with $r_1=1$ and $N=10$. It confirms that the approximations are satisfactory when $r_1 + r_2 < 2\pi$ but erratic outside that domain. We note however that, even there, the errors are bounded.

Truncating the quasi-resolution eq 13 at $n=N$ and $l=L$ yields well-defined approximations to both the operator and its matrix elements. For example, the approximation

$$\tilde{E}^{(N,L)} = \frac{1}{2} \sum_{n=0}^N \sum_{l=0}^L \sum_{m=-l}^l \langle \rho|\phi_{nlm} \rangle \langle \phi_{nlm}|\rho \rangle \quad (22)$$

Table 2. Stewart parameters for atoms

hydrogen		carbon	
c_i	α_i	c_i	α_i
0.29449	0.21	1.71581	0.29
0.63550	0.88	2.54666	0.82
0.05859	3.73	-0.18334	2.31
0.01253	15.90	0.26810	6.50
-0.00111	67.73	1.09048	18.31
		0.45570	51.55
		0.09106	145.16
		0.01337	408.75
		0.00195	1150.99
		0.00016	3241.06
		0.00005	9126.48

has the truncation error

$$E_{TE} = \tilde{E}^{(N,L)} - \tilde{E} \quad (23)$$

Is such a truncation useful in practice? To explore this question, we used eq 22 to calculate the Coulomb self-interaction energy of the electrons in the octahedral nanodiamond $C_{84}H_{64}$ crystallite.²⁷ This molecule has a diamond-like structure with T_d symmetry, and for the sake of simplicity, we have used C–C and C–H bond lengths of 154 and 109 pm, respectively. The electron density

$$\rho(\mathbf{r}) = \sum_{A=1}^{148} \rho_A(\mathbf{r}) \quad (24)$$

is the sum of the Stewart atomic densities^{28–30}

$$\rho_A(\mathbf{r}) = \sum_{i=1}^{D_A} c_i (\alpha_i/\pi)^{3/2} \exp(-\alpha_i |\mathbf{r} - \mathbf{R}_A|^2) \quad (25)$$

generated from the UHF/6-311G densities of isolated 3P carbon and 2S hydrogen atoms. The Stewart parameters are given in Table 2 and yield $E = 20511.5578014$ au.

We have written a C program to compute eq 22, and we use the relative error

$$\varepsilon \equiv \left| \frac{\tilde{E}^{(N,L)} - E}{E} \right| = \left| \frac{E_{DVE} + E_{TE}}{E} \right| \quad (26)$$

to measure the accuracy of the approximation eq 22 for different (N,L) . The molecule's center of mass is placed at the origin, but most of its nuclei still lie outside the allowed domain (i.e., $|\mathbf{R}_A| > \pi$). We therefore compress the entire system by a scale factor s , perform the Coulomb calculation, and then unscale the resulting energy. The relationship between scaled and unscaled systems is described by the following equations

$$\mathbf{R}'_A = s^{-1} \mathbf{R}_A \quad (27)$$

$$\alpha'_i = s^2 \alpha_i \quad (28)$$

$$\rho'(\mathbf{r}') = s^3 \rho(\mathbf{r}) \quad (29)$$

$$E' = sE \quad (30)$$

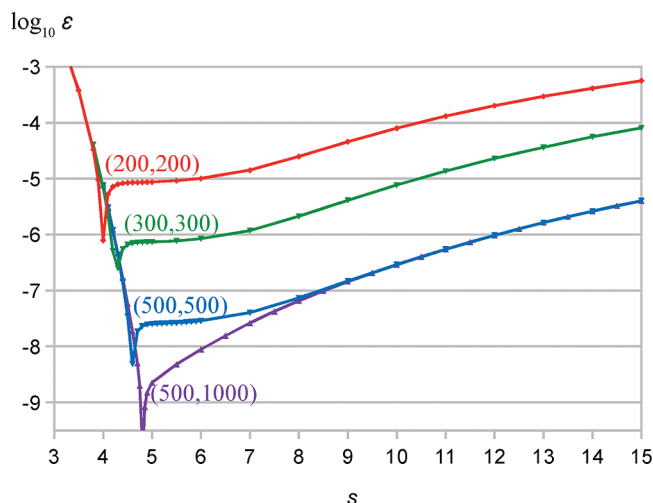


Figure 2. Relative error, eq 26, of $E^{(N,L)}$ for $3 \leq s \leq 15$.

A scaled system described by \mathbf{R}'_i , α'_i , and $\rho'(\mathbf{r}')$ is mathematically equivalent to the unscaled one. Thus, in theory, this scheme is exact and works for any kind of energies or molecular properties. However, when we use scaling in conjunction with truncated resolution, the compression increases the exponents α'_i . As a result, the auxiliary integrals eq 18 decay more slowly, reducing the rate of convergence of eq 22 and increasing the truncation error eq 23.

Figure 2 reveals that there is a DVE-dominated region ($s \lesssim 4$) and a TE-dominated region ($s \gtrsim 5$). The results show that the truncation error grows slowly as s is increased, but that the domain-violation error grows rapidly as s is decreased. It is therefore important to scale the system to fit in the domain, but moderate overcompression does not magnify the error by very much. For $N = 500$ and $L = 1000$, the lowest errors arise near $s = 4.8$, but any s from 4.5 to 12 leads to $\varepsilon < 10^{-6}$.

In summary, we have derived a quasi-resolution of the Coulomb operator that allows it to be expressed in terms of products of one-particle functions. Unlike earlier resolutions, the quasi-resolution is based on simple mathematical functions and is well suited for computational purposes. Our numerical study indicates that the quasi-resolution is useful for computing the Coulomb energy, which is an important bottleneck in DFT calculations. However, the potential scope of the quasi-resolution is much wider than this, and there are significant possibilities for applications to other operators and to exchange and correlation energies. We are currently investigating these and will report results elsewhere.

AUTHOR INFORMATION

Corresponding Author

*E-mail: taweetham.limpanuparb@anu.edu.au.

ACKNOWLEDGMENT

T.L. thanks the Development and Promotion of Science and Technology Talents Project for a Royal Thai Government Ph.D. scholarship. P.M.W.G. thanks the Australian Research Council for funding (DP0984806 and DP1094170). We thank NCI National Facility for a generous allocation of supercomputer resources.

■ REFERENCES

- (1) Gill, P. M. W.; Adamson, R. D. *Chem. Phys. Lett.* **1996**, *261*, 105–110.
- (2) Dombroski, J. P.; Taylor, S. W.; Gill, P. M. W. *J. Phys. Chem.* **1996**, *100*, 6272–6276.
- (3) Adamson, R. D.; Dombroski, J. P.; Gill, P. M. W. *Chem. Phys. Lett.* **1996**, *254*, 329–336.
- (4) Rokhlin, V. J. *Comput. Phys.* **1985**, *60*, 187–207.
- (5) Appel, A. W. *SIAM J. Sci. Stat. Comput.* **1985**, *6*, 85–103.
- (6) Greengard, L. *The rapid evaluation of potential fields in particle systems*; MIT Press: Cambridge, MA, 1987.
- (7) White, C. A.; Johnson, B. G.; Gill, P. M. W.; Head-Gordon, M. *Chem. Phys. Lett.* **1996**, *253*, 268–278.
- (8) Fusti-Molnar, L.; Pulay, P. *J. Chem. Phys.* **2002**, *117*, 7827–7835.
- (9) Beebe, N. H. F.; Linderberg, J. *Int. J. Quantum Chem.* **1977**, *12*, 683–705.
- (10) Koch, H.; Sanchez de Meras, A.; Pedersen, T. B. *J. Chem. Phys.* **2003**, *118*, 9481–9484.
- (11) Aquilante, F.; Lindh, R.; Pedersen, T. B. *J. Chem. Phys.* **2007**, *127*, 114107.
- (12) Weigend, F.; Kattannek, M.; Ahlrichs, R. *J. Chem. Phys.* **2009**, *130*, 164106.
- (13) Vahtras, O.; Almlöf, J.; Feyereisen, M. *Chem. Phys. Lett.* **1993**, *213*, 514–518.
- (14) Jung, Y.; Sodt, A.; Gill, P. M. W.; Head-Gordon, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6692–6697.
- (15) Chinnamsetty, S. R.; Espig, M.; Khoromskij, B. N.; Hackbusch, W.; Flad, H.-J. *J. Chem. Phys.* **2007**, *127*, 084110.
- (16) Kinoshita, T.; Hino, O.; Bartlett, R. J. *J. Chem. Phys.* **2003**, *119*, 7756–7762.
- (17) Hino, O.; Kinoshita, T.; Bartlett, R. J. *J. Chem. Phys.* **2004**, *121*, 1206–1213.
- (18) Varganov, S. A.; Gilbert, A. T. B.; Deplazes, E.; Gill, P. M. W. *J. Chem. Phys.* **2008**, *128*, 201104.
- (19) Gill, P. M. W.; Gilbert, A. T. B. *Chem. Phys.* **2009**, *356*, 86–90.
- (20) Limpanuparb, T.; Gill, P. M. W. *Phys. Chem. Chem. Phys.* **2009**, *11*, 9176–9181.
- (21) Olver, F. W. J.; Lozier, D. W.; Boisvert, R. F.; Clark, C. W. *NIST Handbook of Mathematical Functions*; Cambridge University Press: New York, 2010.
- (22) Dominici, D. E.; Gill, P. M. W.; Limpanuparb, T. A Remarkable Identity Involving Bessel Functions, arXiv:1103.0058. arXiv.org ePrint archive. <http://arxiv.org/abs/1103.0058> (accessed Mar 1, 2011).
- (23) Gradshteyn, I. S.; Ryzhik, I. M. *Table of Integrals, Series and Products*; Academic: London, 2007; pp 683–684.
- (24) R: *A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2010.
- (25) Smith, J. M.; Olver, F. W. J.; Lozier, D. W. *ACM Trans. Math. Software* **1981**, *7*, 93–105.
- (26) Boyd, J. P. *Chebyshev and Fourier Spectral Methods*, 2nd ed.; Dover: New York, 2000; pp 32–35.
- (27) Filik, J.; Harvey, J. N.; Allan, N. L.; May, P. W.; Dahl, J. E. P.; Liu, S.; Carlson, R. M. K. *Phys. Rev. B* **2006**, *74*, 035423.
- (28) Gill, P. M. W. *J. Phys. Chem.* **1996**, *100*, 15421–15427.
- (29) Lee, A. M.; Gill, P. M. W. *Chem. Phys. Lett.* **1998**, *286*, 226–232.
- (30) Gilbert, A. T. B.; Gill, P. M. W.; Taylor, S. W. *J. Chem. Phys.* **2004**, *120*, 7887–7893.

■ NOTE ADDED AFTER ASAP PUBLICATION

This Letter was published ASAP on March 16, 2011. A correction has been made to equation 26. The correct version was published on March 21, 2011.

The Langevin Hull: Constant Pressure and Temperature Dynamics for Nonperiodic Systems

Charles F. Vardeman, II, Kelsey M. Stocker, and J. Daniel Gezelter*

Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, Indiana 46556, United States

ABSTRACT: We have developed a new isobaric–isothermal (NPT) algorithm which applies an external pressure to the facets comprising the convex hull surrounding the system. A Langevin thermostat is also applied to the facets to mimic contact with an external heat bath. This new method, the “Langevin Hull”, can handle heterogeneous mixtures of materials with different compressibilities. These systems are problematic for traditional affine transform methods. The Langevin Hull does not suffer from the edge effects of boundary potential methods and allows realistic treatment of both external pressure and thermal conductivity due to the presence of an implicit solvent. We apply this method to several different systems, including bare metal nanoparticles and nanoparticles in an explicit solvent as well as clusters of liquid water. The predicted mechanical properties of these systems are in good agreement with experimental data and previous simulation work.

1. INTRODUCTION

The most common molecular dynamics methods for sampling configurations from an isobaric–isothermal (NPT) ensemble maintain a target pressure in a simulation by coupling the volume of the system to a barostat, which is an extra degree of freedom propagated along with the particle coordinates. These methods require periodic boundary conditions because when the instantaneous pressure in the system differs from the target pressure, the volume is reduced or expanded using affine transforms of the system geometry. An affine transform scales the size and the shape of the periodic box as well as the particle positions within the box (but not the sizes of the particles). The most common constant pressure methods, including the Melchionna modification¹ to the Nosé–Hoover–Andersen equations of motion,^{2–4} the Berendsen pressure bath,⁵ and the Langevin Piston,^{6,7} all utilize scaled coordinate transformation to adjust the box volume. As long as the material in the simulation box has a relatively uniform compressibility, the standard affine transform approach provides an excellent way of adjusting the volume of the system and applying pressure directly via the interactions between atomic sites.

One problem with this approach appears when the system being simulated is an inhomogeneous mixture in which portions of the simulation box are incompressible relative to other portions. Examples include simulations of metallic nanoparticles in liquid environments and proteins at ice/water interfaces as well as other heterogeneous or interfacial environments. In these cases, the affine transform of atomic coordinates either will cause numerical instability when the sites in the incompressible medium collide with each other or will lead to inefficient sampling of system volumes if the barostat is set slow enough to avoid the instabilities in the incompressible region.

One may also wish to avoid affine transform periodic boundary methods to simulate explicitly nonperiodic systems under constant pressure conditions. The use of periodic boxes to enforce a system volume requires either effective solute concentrations that are much higher than desirable or unreasonable

system sizes to avoid this effect. For example, calculations using typical hydration boxes solvating a protein under periodic boundary conditions are quite expensive. A 62 Å³ box of water solvating a moderately small protein, like hen egg white lysozyme (PDB code: 1LYZ), yields an effective protein concentration of 100 mg/mL.⁸

Total protein concentrations in the cell are typically on the order of 160–310 mg/mL,⁹ and individual proteins have concentrations orders of magnitude lower than this in the cellular environment. The effective concentrations of single proteins in simulations may have significant effects on the structure and dynamics of simulated systems.

1.1. Boundary Methods. There have been a number of approaches to handle simulations of explicitly nonperiodic systems that focus on constant or nearly constant volume conditions while maintaining bulk-like behavior. Berkowitz and McCammon introduced a stochastic (Langevin) boundary layer inside a region of fixed molecules which effectively enforces constant temperature and volume (NVT) conditions.¹⁰ In this approach, the stochastic and fixed regions were defined relative to a central atom. Brooks and Karplus extended this method to include deformable stochastic boundaries.¹¹ The stochastic boundary approach has been used widely for protein simulations.

The electrostatic and dispersive behavior near the boundary has long been a cause for concern when performing simulations of explicitly nonperiodic systems. Early work led to the surface constrained soft sphere dipole model (SCSSD)¹² in which the surface molecules are fixed in a random orientation representative of the bulk solvent structural properties. Belch et al.¹³ simulated clusters of TIPS2 water surrounded by a hydrophobic bounding potential. The spherical hydrophobic boundary induced dangling hydrogen bonds at the surface that propagated deep into the cluster, affecting most of the molecules in the simulation. This result echoes an earlier study which showed that

Received: November 18, 2010

Published: March 18, 2011

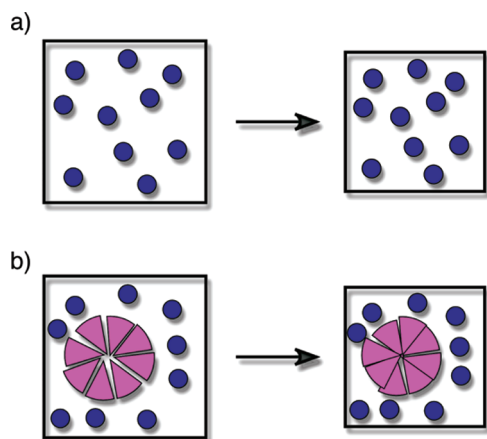


Figure 1. Affine scaling methods use box-length scaling to adjust the volume to under- or overpressure conditions. In a system with a uniform compressibility (e.g., bulk fluids), these methods can work well. In systems containing heterogeneous mixtures, the affine scaling moves required to adjust the pressure in the high-compressibility regions can cause molecules in low-compressibility regions to collide.

an extended planar hydrophobic surface caused orientational preferences at the surface which extended relatively deep (7 Å) into the liquid simulation cell.¹⁴ The surface constrained all-atom solvent (SCAAS) model¹⁵ improved upon its SCSSD predecessor. The SCAAS model utilizes a polarization constraint which is applied to the surface molecules to maintain bulk-like structure at the cluster surface. A radial constraint is used to maintain the desired bulk density of the liquid. Both constraint forces are applied only to a predetermined number of the outermost molecules.

Beglov and Roux have developed a boundary model in which the hard sphere boundary has a radius that varies with the instantaneous configuration of the solute (and solvent) molecules.¹⁶ This model contains a clear pressure and a surface tension contribution to the free energy.

1.2. Restraining Potentials. Restraining potentials introduce repulsive potentials at the surface of a sphere or other geometry. The solute and any explicit solvent are therefore restrained inside the range defined by the external potential. Often the potentials include a weak short-range attraction to maintain the correct density at the boundary. Beglov and Roux have also introduced a restraining boundary potential which relaxes dynamically depending on the solute geometry and the force the explicit system exerts on the shell.¹⁷

Recently, Krilov et al. introduced a flexible boundary model that uses a Lennard-Jones potential between the solvent molecules and a boundary which is determined dynamically from the position of the nearest solute atom.^{18,19} This approach allows the confining potential to prevent solvent molecules from migrating too far from the solute surface, while providing a weak attractive force pulling the solvent molecules toward a fictitious bulk solvent. Although this approach is appealing and has physical motivation, nanoparticles do not deform far from their original geometries even at temperatures which vaporize the nearby solvent. For the systems like this, the flexible boundary model will be nearly identical to a fixed volume restraining potential.

1.3. Hull Methods. The approach of Kohanoff, Caro, and Finnis is the most promising of the methods for introducing both constant pressure and temperature into nonperiodic simulations.^{20,21} This method is based on standard Langevin dynamics, but the

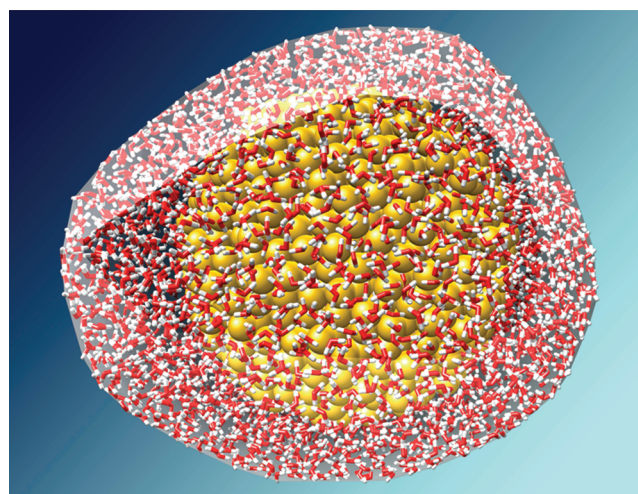


Figure 2. The external temperature and pressure bath interacts only with those atoms on the convex hull (gray surface). The hull is computed dynamically at each time step, and molecules can move between the interior (Newtonian) region and the Langevin Hull.

Brownian or random forces are allowed to act only on peripheral atoms and exert forces in a direction that is inward-facing relative to the facets of a closed bounding surface. The statistical distribution of the random forces are uniquely tied to the pressure in the external reservoir, so the method can be shown to sample the isobaric–isothermal ensemble. Kohanoff et al. used a Delaunay tessellation to generate a bounding surface surrounding the outermost atoms in the simulated system. This is not the only possible triangulated outer surface but guarantees that all of the random forces point inward toward the cluster.

In the following sections, we extend and generalize the approach of Kohanoff, Caro, and Finnis. The new method, which we are calling the “Langevin Hull” applies the external pressure, Langevin drag, and random forces on the facets of the hull instead of the atomic sites comprising the vertices of the hull. This allows us to decouple the external pressure contribution from the drag and random force. The methodology is introduced in Section 2, tests on crystalline nanoparticles, liquid clusters, and heterogeneous mixtures are detailed in Section 3. Section 4 summarizes our findings.

2. METHODOLOGY

The Langevin Hull uses an external bath at a fixed constant pressure (P) and temperature (T) with an effective solvent viscosity (η). This bath interacts only with the objects on the exterior hull of the system. Defining the hull of the atoms in a simulation is done in a manner similar to the approach of Kohanoff, Caro, and Finnis.²⁰ That is, any instantaneous configuration of the atoms in the system is considered as a point cloud in three-dimensional space. Delaunay triangulation is used to finding all facets between coplanar neighbors.^{22,23} In highly symmetric point clouds, facets can contain many atoms, but in all but the most symmetric of cases, the facets are simple triangles in three-space which contain exactly three atoms.

The convex hull is the set of facets that have no concave corners at an atomic site.^{24,25} This eliminates all facets on the interior of the point cloud, leaving only those exposed to the bath. Sites on the convex hull are dynamic; as molecules re-enter

the cluster, all interactions between atoms on that molecule and the external bath are removed. Since the edge is determined dynamically as the simulation progresses, no a priori geometry is defined. The pressure and temperature bath interacts only with the atoms on the edge and not with atoms interior to the simulation.

Atomic sites in the interior of the simulation move under standard Newtonian dynamics:

$$m_i \dot{\mathbf{v}}_i(t) = -\nabla_i U \quad (1)$$

where m_i is the mass of site i , $\mathbf{v}_i(t)$ is the instantaneous velocity of site i at time t , and U is the total potential energy. For atoms on the exterior of the cluster (i.e., those that occupy one of the vertices of the convex hull), the equation of motion is modified with an external force, $\mathbf{F}_i^{\text{ext}}$:

$$m_i \dot{\mathbf{v}}_i(t) = -\nabla_i U + \mathbf{F}_i^{\text{ext}} \quad (2)$$

The external bath interacts indirectly with the atomic sites through the intermediary of the hull facets. Since each vertex (or atom) provides one corner of a triangular facet, the force on the facets are divided equally to each vertex. However, each vertex can participate in multiple facets, so the resultant force is a sum over all facets f containing vertex i :

$$\mathbf{F}_i^{\text{ext}} = \sum_{\text{facets } f \text{ containing } i} \frac{1}{3} \mathbf{F}_f^{\text{ext}} \quad (3)$$

The external pressure bath applies a force to the facets of the convex hull in direct proportion to the area of the facet, while the thermal coupling depends on the solvent temperature, viscosity, and the size and shape of each facet. The thermal interactions are expressed as a standard Langevin description of the forces:

$$\begin{aligned} \mathbf{F}_f^{\text{ext}} &= \text{external pressure} + \text{drag force} + \text{random force} \\ &= -\hat{n}_f P A_f - \Xi_f(t) \mathbf{v}_f(t) + \mathbf{R}_f(t) \end{aligned} \quad (4)$$

Here, A_f and \hat{n}_f are the area and (outward-facing) normal vectors for facet f , respectively. While $\mathbf{v}_f(t)$ is the velocity of the facet centroid:

$$\mathbf{v}_f(t) = \frac{1}{3} \sum_{i=1}^3 \mathbf{v}_i \quad (5)$$

and $\Xi_f(t)$ is an approximate (3×3) resistance tensor that depends on the geometry and surface area of facet f and the viscosity of the bath. The resistance tensor is related to the fluctuations of the random force, $\mathbf{R}(t)$, by the fluctuation–dissipation theorem:

$$\langle \mathbf{R}_f(t) \rangle = 0 \quad (6)$$

$$\langle \mathbf{R}_f(t) \mathbf{R}_f^T(t') \rangle = 2k_B T \Xi_f(t) \delta(t - t') \quad (7)$$

Once the resistance tensor is known for a given facet, a stochastic vector that has the properties in eq 7 can be calculated efficiently by carrying out a Cholesky decomposition to obtain the square root matrix of the resistance tensor:

$$\Xi_f = \mathbf{S} \mathbf{S}^T \quad (8)$$

where \mathbf{S} is a lower triangular matrix.²⁶ A vector with the statistics required for the random force can then be obtained by multiplying \mathbf{S} onto a random three-vector \mathbf{Z} , which has elements

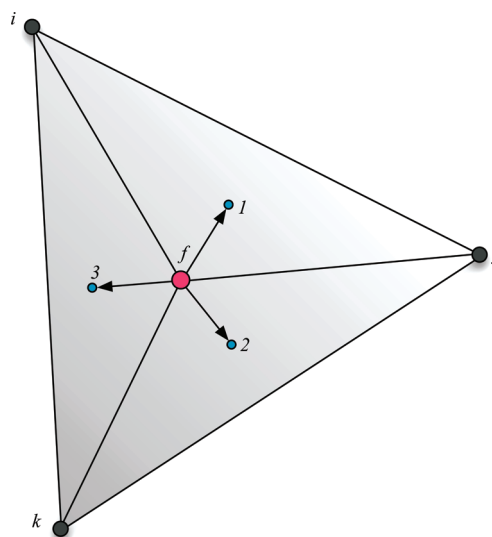


Figure 3. The resistance tensor Ξ for a facet comprising sites i, j , and k is constructed using Oseen tensor contributions between the centroid of the facet f and each of the subfacets ((if, j) , (jf, k) , and (kf, i)). The centroids of the subfacets are located at 1, 2, and 3, and the area of each subfacet is easily computed using half of the cross product of two of the edges.

chosen from a Gaussian distribution, such that

$$\langle \mathbf{Z}_i \rangle = 0, \quad \langle \mathbf{Z}_i \cdot \mathbf{Z}_j \rangle = \frac{2k_B T}{\delta t} \delta_{ij} \quad (9)$$

where δt is the time step in use during the simulation. The random force, $\mathbf{R}_f = \mathbf{S} \mathbf{Z}$, can be shown to have the correct properties required by eq 7.

Our treatment of the resistance tensor is approximate. Ξ_f for a rigid triangular plate would normally be treated as a 6×6 tensor that includes translational and rotational drag as well as translational–rotational coupling. The computation of resistance tensors for rigid bodies has been detailed elsewhere,^{27–30} but the standard approach involving bead approximations would be prohibitively expensive if it were recomputed at each step in a molecular dynamics simulation.

Instead, we are utilizing an approximate resistance tensor obtained by first constructing the Oseen tensor for the interaction of the centroid of the facet (f) with each of the subfacets $\ell = 1, 2, 3$,

$$T_{\ell f} = \frac{A_\ell}{8\pi\eta R_{\ell f}} \left(I + \frac{\mathbf{R}_{\ell f} \mathbf{R}_{\ell f}^T}{R_{\ell f}^2} \right) \quad (10)$$

Here, A_ℓ is the area of subfacet ℓ which is a triangle containing two of the vertices of the facet along with the centroid. $\mathbf{R}_{\ell f}$ is the vector between the centroid of facet f and the centroid of subfacet ℓ , I is the (3×3) identity matrix, and η is the viscosity of the external bath.

The tensors for each of the subfacets are added together, and the resulting matrix is inverted to give a 3×3 resistance tensor for translations of the triangular facet:

$$\Xi_f(t) = \left[\sum_{i=1}^3 T_{if} \right]^{-1} \quad (11)$$

Note that this treatment ignores rotations (and translational–rotational coupling) of the facet. In compact systems, the

facets stay relatively fixed in orientation between configurations, so this appears to be a reasonably good approximation.

We have implemented this method by extending the Langevin dynamics integrator in our code, OpenMD.^{31,32} At each molecular dynamics time step, the following process is carried out:

- (1) The standard interatomic forces ($\nabla_i U$) are computed.
- (2) Delaunay triangulation is carried out using the current atomic configuration.
- (3) The convex hull is computed and facets are identified.
- (4) For each facet:
 - (a) The force from the pressure bath ($-\hat{n}_i P A_i$) is computed.
 - (b) The resistance tensor [$\Xi_f(t)$] is computed using the viscosity (η) of the bath.
 - (c) Facet drag [$-\Xi_f(t)v_f(t)$] forces are computed.
 - (d) Random forces [$R_f(t)$] are computed using the resistance tensor and the temperature (T) of the bath.
- (5) The facet forces are divided equally among the vertex atoms.
- (6) Atomic positions and velocities are propagated.

The Delaunay triangulation and computation of the convex hull are done using calls to the qhull library.³³ There is a minimal penalty for computing the convex hull and the resistance tensors at each step in the molecular dynamics simulation (roughly $0.02 \times$ cost of a single force evaluation), and the convex hull is remarkably easy to parallelize on distributed memory machines (see Appendix A).

3. TESTS AND APPLICATIONS

To test the new method, we have carried out simulations using the Langevin Hull on: (1) a crystalline system (gold nanoparticles), (2) a liquid droplet (SPC/E water),³⁴ and (3) a heterogeneous mixture (gold nanoparticles in an SPC/E water droplet). In each case, we have computed properties that depend on the external applied pressure. Of particular interest for the single-phase systems is the isothermal compressibility:

$$\kappa_T = -\frac{1}{V} \left(\frac{\partial V}{\partial P} \right)_T \quad (12)$$

One problem with eliminating periodic boundary conditions and simulation boxes is that the volume of a three-dimensional point cloud is not well-defined. In order to compute the compressibility of a bulk material, we make an assumption that the number density, $\rho = N/V$, is uniform within some region of the point cloud. The compressibility can then be expressed in terms of the average number of particles in that region:

$$\kappa_T = -\frac{1}{N} \left(\frac{\partial N}{\partial P} \right)_T \quad (13)$$

The region we used is a spherical volume of 20 Å radius centered in the middle of the cluster with a roughly 25 Å radius. N is the average number of molecules found within this region throughout a given simulation. The geometry of the region is arbitrary, and any bulk-like portion of the cluster can be used to compute the compressibility.

One might assume that the volume of the convex hull could simply be taken as the system volume V in the compressibility expression (eq 12), but this has implications at lower pressures (which are explored in detail in the section on water droplets).

The metallic force field in use for the gold nanoparticles is the quantum Sutton–Chen (QSC) model.³⁵ In all simulations involving point charges, we utilized damped shifted force (DSF) electrostatics,³⁶ which is a variant of the Wolf summation³⁷ that has been shown to provide good forces and torques on molecular models for water in a computationally efficient manner.³⁶ The damping parameter (α) was set to 0.18 \AA^{-1} , and the cutoff radius was set to 12 Å. The Spohr potential was adopted in depicting the interaction between metal atoms and the SPC/E water molecules.³⁸

3.1. Bulk Modulus of Gold Nanoparticles. The compressibility (and its inverse, the bulk modulus) is well-known for gold and is captured well by the embedded atom method (EAM)³⁹ potential and related many-body force fields. In particular, the quantum Sutton–Chen (SC) potential gets nearly quantitative agreement with the experimental bulk modulus values and makes a good first test of how the Langevin Hull will perform at large applied pressures.

The SC potentials are based on a model of a metal which treats the nuclei and the core electrons as pseudoatoms embedded in the electron density due to the valence electrons on all of the other atoms in the system.⁴⁰ The SC potential has a simple form that closely resembles the Lennard-Jones potential:

$$U_{tot} = \sum_i \left[\frac{1}{2} \sum_{j \neq i} D_{ij} V_{ij}^{\text{pair}}(r_{ij}) - c_i D_{ii} \sqrt{\rho_i} \right] \quad (14)$$

where V_{ij}^{pair} and ρ_i are given by

$$V_{ij}^{\text{pair}}(r) = \left(\frac{\alpha_{ij}}{r_{ij}} \right)^{n_{ij}}, \quad \rho_i = \sum_{j \neq i} \left(\frac{\alpha_{ij}}{r_{ij}} \right)^{m_{ij}} \quad (15)$$

V_{ij}^{pair} is a repulsive pairwise potential that accounts for interactions between the pseudoatom cores. The $(\rho_i)^{1/2}$ term in eq 14 is an attractive many-body potential that models the interactions between the valence electrons and the cores of the pseudoatoms. D_{ij} and D_{ii} set the appropriate overall energy scale, c_i scales the attractive portion of the potential relative to the repulsive interaction, and α_{ij} is a length parameter that assures a dimensionless form for ρ . These parameters are tuned to various experimental properties, such as the density, cohesive energy, and elastic moduli for FCC transition metals. The QSC formulation matches these properties while including zero-point quantum corrections for different transition metals.^{35,41}

In bulk gold, the experimentally measured value for the bulk modulus is 180.32 GPa, while previous calculations on the QSC potential in periodic boundary simulations of the bulk crystal have yielded values of 175.53 GPa.⁴¹ Using the same force field, we have performed a series of 1 ns simulations on gold nanoparticles of three different radii: 20 Å (1985 atoms), 30 Å (6699 atoms), and 40 Å (15707 atoms) utilizing the Langevin Hull at a variety of applied pressures ranging from 0 to 10 GPa. For the 40 Å radius nanoparticle we obtain a value of 177.55 GPa for the bulk modulus of gold, in close agreement with both previous simulations and the experimental bulk modulus reported for gold single crystals.⁴² The smaller gold nanoparticles (30 and 20 Å radii) have calculated bulk moduli of 215.58 and 208.86 GPa, respectively, indicating that smaller nanoparticles are somewhat stiffer (less compressible) than the larger nanoparticles. This stiffening of the small nanoparticles may be related to their high degree of surface curvature, resulting in a lower

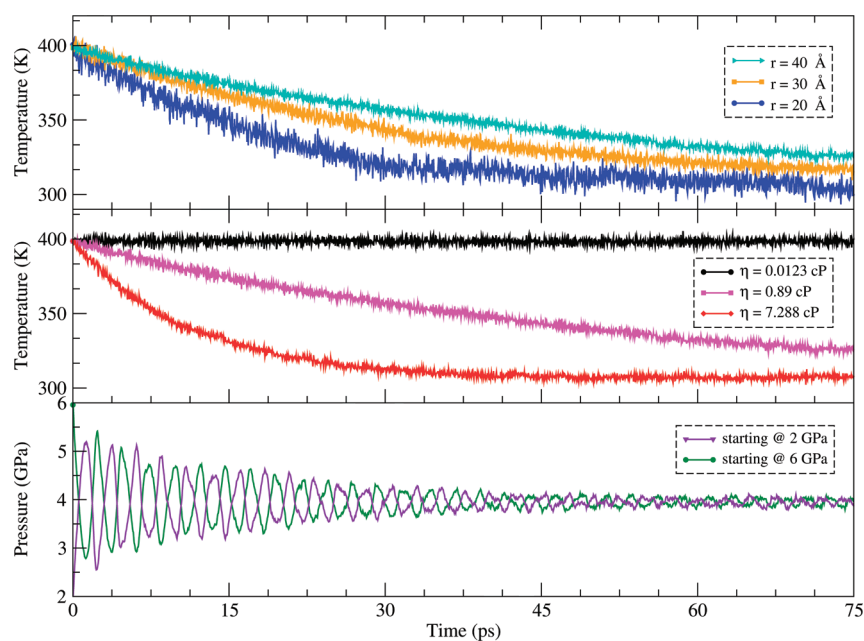


Figure 4. The response of the internal pressure and temperature of gold nanoparticles when first placed in the Langevin Hull ($T_{\text{bath}} = 300$ K, $P_{\text{bath}} = 4$ GPa), starting from initial conditions that were far from the bath pressure and temperature. The pressure response is rapid (after the breathing mode oscillations in the nanoparticle die out), and the rate of thermal equilibration depends on both the exposed surface area (top panel) and the viscosity of the bath (middle panel).

coordination number of surface atoms relative to the surface atoms in the 40 Å radius particle.

We obtain a gold lattice constant of 4.051 Å using the Langevin Hull at 1 atm, close to the experimentally determined value for bulk gold and the value for gold simulated using the QSC potential and periodic boundary conditions (4.079 Å and 4.088 Å, respectively).⁴¹ The slightly smaller calculated lattice constant is most likely due to the presence of surface tension in the nonperiodic Langevin Hull cluster, an effect absent from a bulk simulation. The specific heat of a 40 Å gold nanoparticle under the Langevin Hull at 1 atm is 24.914 J/mol K, which compares very well with the experimental value of 25.42 J/mol K.

We note that the Langevin Hull produces rapidly converging behavior for structures that are started far from equilibrium. In Figure 4 we show how the pressure and temperature respond to the Langevin Hull for nanoparticles that were initialized far from the target pressure and temperature. As expected, the rate at which thermal equilibrium is achieved depends on the total surface area of the cluster exposed to the bath as well as the bath viscosity. Pressure that is applied suddenly to a cluster can excite breathing vibrations, but these rapidly damp out (on time scales of 30–50 ps).

3.2. Compressibility of SPC/E Water Clusters. Prior molecular dynamics simulations on SPC/E water (both in NVT⁴³ and NPT^{44,45} ensembles) have yielded values for the isothermal compressibility that agree well with experiment.⁴⁶ The results of two different approaches for computing the isothermal compressibility from Langevin Hull simulations for pressures between 1 and 3000 atm are shown in Figure 5 along with compressibility values obtained from both other SPC/E simulations and experiment.

Isothermal compressibility values calculated using the number density (eq 13) expression are in good agreement with experimental and previous simulation work throughout the 1–1000

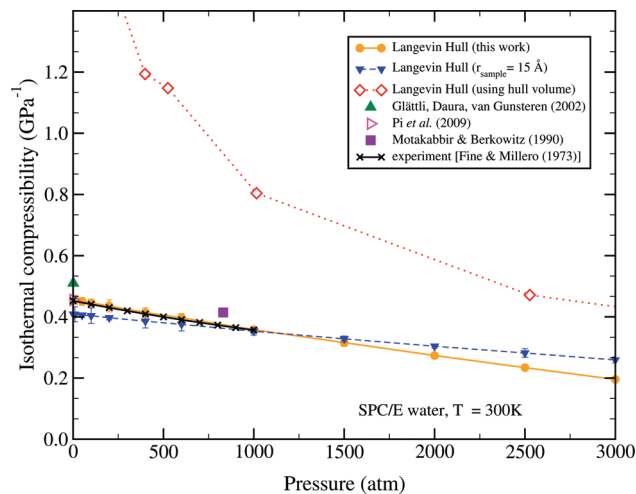


Figure 5. Compressibility of SPC/E water.

atm pressure regime. Compressibilities computed using the Hull volume, however, deviate dramatically from the experimental values at low applied pressures. The reason for this deviation is quite simple: At low applied pressures, the liquid is in equilibrium with a vapor phase, and it is entirely possible for one or a few molecules to drift away from the liquid cluster (see Figure 6). At low pressures, the restoring forces on the facets are very gentle, and this means that the hulls often take on relatively distorted geometries, which include large volumes of empty space.

At higher pressures, the equilibrium strongly favors the liquid phase, and the hull geometries are much more compact. Because of the liquid–vapor effect on the convex hull, the regional number density approach (eq 13) provides more reliable estimates of the compressibility.

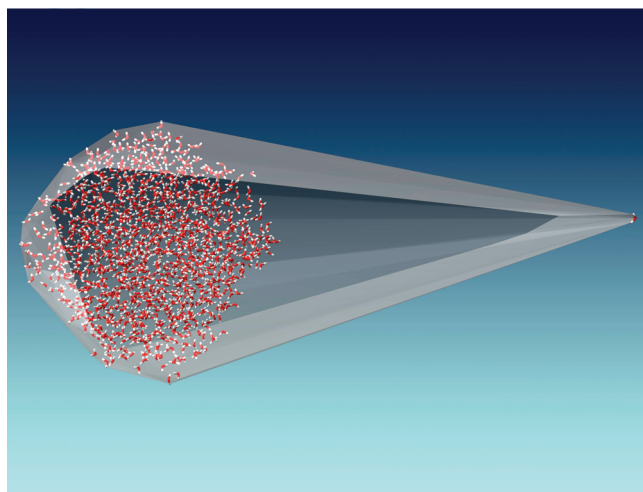


Figure 6. At low pressures, the liquid is in equilibrium with the vapor phase, and isolated molecules can detach from the liquid droplet. This is expected behavior, but the volume of the convex hull includes large regions of empty space. For this reason, compressibilities are computed using local number densities rather than hull volumes.

In both the traditional compressibility formula (eq 12) and the number density version (eq 13), multiple simulations at different pressures must be done to compute the first derivatives. It is also possible to compute the compressibility using the fluctuation dissipation theorem using either fluctuations in the volume:⁴⁷

$$\kappa_T = \frac{\langle V^2 \rangle - \langle V \rangle^2}{Vk_B T} \quad (16)$$

or, equivalently, fluctuations in the number of molecules within the fixed region:

$$\kappa_T = \frac{\langle N^2 \rangle - \langle N \rangle^2}{Nk_B T} \quad (17)$$

Thus, the compressibility of each simulation can be calculated entirely independently from other trajectories. Compressibility calculations that rely on the hull volume will still suffer the effects of the empty space due to the vapor phase. For this reason, we recommend using the number density (eq 13) or number density fluctuations (eq 17) for computing compressibilities. We obtained the results in Figure 5 using a sampling radius that was approximately 80% of the mean distance between the center of mass of the cluster and the hull atoms. This ratio of sampling radius to average hull radius excludes the problematic vapor phase on the outside of the cluster, while including enough of the liquid phase to avoid poor statistics due to fluctuating local densities.

A comparison of the oxygen–oxygen radial distribution functions for SPC/E water simulated using both the Langevin Hull and the more traditional periodic boundary methods—both at 1 atm and 300K—reveals an understructuring of water in the Langevin Hull that manifests as a slight broadening of the solvation shells. This effect may be due to the introduction of a liquid–vapor interface in the Langevin Hull simulations (an interface which is missing in most periodic simulations of bulk water). Vapor-phase molecules contribute a small but nearly flat portion of the radial distribution function.

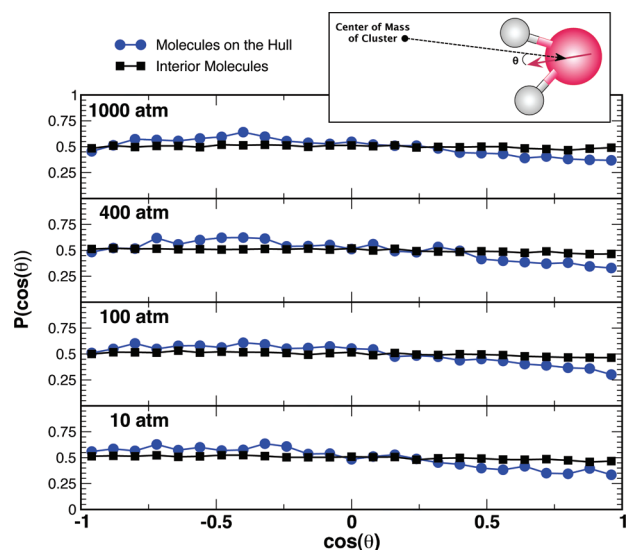


Figure 7. Distribution of $\cos \theta$ values for molecules on the interior of the cluster (squares) and for those participating in the convex hull (circles) at a variety of pressures. The Langevin Hull exhibits minor dewetting behavior with exposed oxygen sites on the hull water molecules. The orientational preference for exposed oxygen appears to be independent of applied pressure.

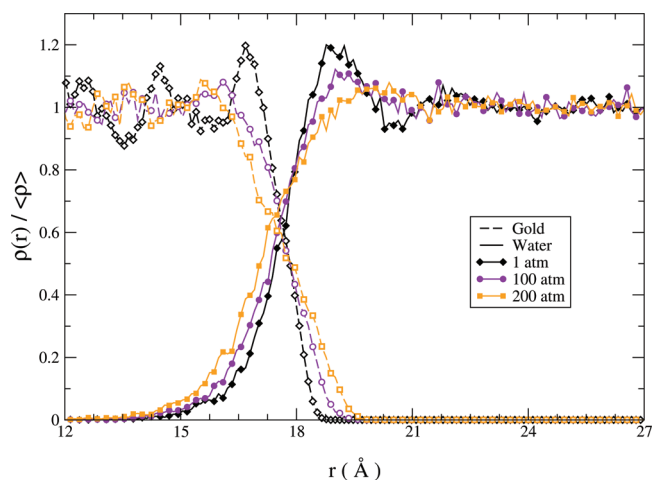


Figure 8. Density profiles of gold and water at the nanoparticle surface. Each curve has been normalized by the average density in the bulk-like region available to the corresponding material. Higher applied pressures destructure both the gold nanoparticle surface and the water at the metal/water interface.

3.3. Molecular Orientation Distribution at Cluster Boundary. In order for a nonperiodic boundary method to be widely applicable, it must be constructed in such a way that they allow a finite system to replicate the properties of the bulk. Early nonperiodic simulation methods (e.g., hydrophobic boundary potentials) induced spurious orientational correlations deep within the simulated system.^{13,14} This behavior spawned many methods for fixing and characterizing the effects of artificial boundaries, including methods which fix the orientations of a set of edge molecules.^{12,15}

As described above, the Langevin Hull does not require that the orientation of molecules be fixed nor does it utilize an

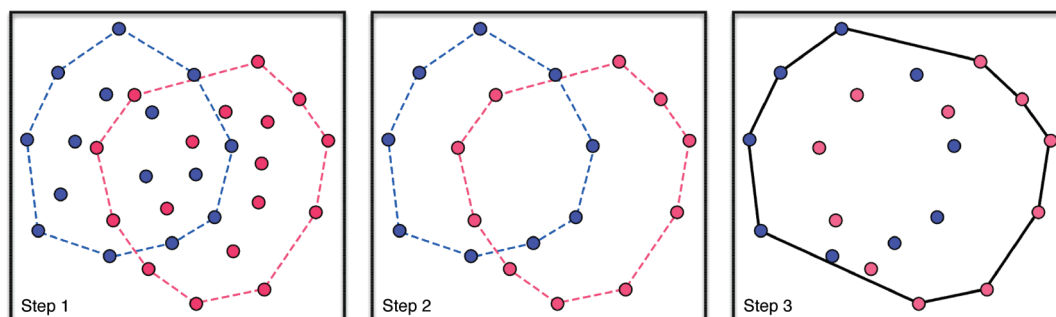


Figure 9. When the sites are distributed among many nodes for parallel computation, the processors first compute the convex hulls for their own sites (dashed lines in left panel). The positions of the sites that make up the subset hulls are then communicated to all processors (middle panel). The convex hull of the system (solid line in right panel) is the convex hull of the points on the union of the subset hulls.

explicitly hydrophobic boundary or orientational or radial constraints. Therefore, the orientational correlations of the molecules in water clusters are of particular interest in testing this method. Ideally, the water molecules on the surfaces of the clusters will have enough mobility into and out of the center of the cluster to maintain bulk-like orientational distribution in the absence of orientational and radial constraints. However, since the number of hydrogen-bonding partners available to molecules on the exterior is limited, it is likely that there will be an effective hydrophobicity of the hull.

To determine the extent of these effects, we examined the orientations exhibited by SPC/E water in a cluster of 1372 molecules at 300 K and at pressures ranging from 1 to 1000 atm. The orientational angle of a water molecule is described by

$$\cos \theta = \frac{\vec{r}_i \cdot \vec{\mu}_i}{|\vec{r}_i| |\mu_i|} \quad (18)$$

where \vec{r}_i is the vector between molecule i 's center of mass and the cluster center of mass, and $\vec{\mu}_i$ is the vector bisecting the H–O–H angle of molecule i . Bulk-like distributions will result in $\langle \cos \theta \rangle$ values close to zero. If the hull exhibits an overabundance of externally oriented oxygen sites, then the average orientation will be negative, while dangling hydrogen sites will result in positive average orientations.

Figure 7 shows the distribution of $\cos \theta$ values for molecules in the interior of the cluster (squares) and for molecules included in the convex hull (circles).

As expected, interior molecules (those not included in the convex hull) maintain a bulk-like structure with a uniform distribution of orientations. Molecules included in the convex hull show a slight preference for values of $\cos \theta < 0$. These values correspond to molecules with oxygen directed toward the exterior of the cluster, forming dangling hydrogen-bond acceptor sites.

The orientational preference exhibited by water molecules on the hull is significantly weaker than the preference caused by an explicit hydrophobic bounding potential. Additionally, the Langevin Hull does not require that the orientation of any molecules be fixed in order to maintain bulk-like structure, even near the cluster surface.

Previous molecular dynamics simulations of SPC/E liquid/vapor interfaces using periodic boundary conditions have shown that molecules on the liquid side of the interface favor a similar orientation, where oxygen is directed away from the bulk.⁴⁸ These simulations had well-defined liquid- and vapor-phase

regions equilibrium, and it was observed that vapor molecules generally had one hydrogen protruding from the surface, forming a dangling hydrogen-bond donor. Our water clusters do not have a true vapor region but rather a few transient molecules that leave the liquid droplet (and which return to the droplet relatively quickly). Although we cannot obtain an orientational preference of vapor-phase molecules in a Langevin Hull simulation, we do agree with previous estimates of the orientation of liquid-phase molecules at the interface.

3.4. Heterogeneous Nanoparticle/Water Mixtures. To further test the method, we simulated gold nanoparticles ($r = 18 \text{ \AA}$, 1433 atoms) solvated by explicit SPC/E water clusters (5000 molecules) using a model for the gold/water interactions that has been used by Dou et al. for investigating the separation of water films near hot metal surfaces.³⁸ The Langevin Hull was used to sample pressures of 1, 2, 5, 10, 20, 50, 100, and 200 atm, while all simulations were done at a temperature of 300 K. At these temperatures and pressures, there is no observed separation of the water film from the surface.

In Figure 8 we show the density of water and gold as a function of the distance from the center of the nanoparticle. Higher applied pressures appear to destroy structural correlations in the outermost monolayer of the gold nanoparticle as well as in the water near the metal/water interface. Simulations at increased pressures exhibit significant overlap of the gold and water densities, indicating a less well-defined interfacial surface.

At even higher pressures (500 atm and above), problems with the metal–water interaction potential became quite clear. The model we are using appears to have been parametrized for relatively low pressures; it utilizes both shifted Morse and repulsive Morse potentials to model the Au/O and Au/H interactions, respectively. The repulsive wall of the Morse potential does not diverge quickly enough at short distances to prevent water from diffusing into the center of the gold nanoparticles. This behavior is likely not a realistic description of the real physics of the situation. A better model of the gold–water adsorption behavior would require harder repulsive walls to prevent this behavior.

4. DISCUSSION

The Langevin Hull samples the isobaric–isothermal ensemble for nonperiodic systems by coupling the system to a bath characterized by pressure, temperature, and solvent viscosity. This enables the simulation of heterogeneous systems composed of materials with significantly different compressibilities. Because

the boundary is dynamically determined during the simulation and the molecules interacting with the boundary can change, the method inflicts minimal perturbations on the behavior of molecules at the edges of the simulation. Further work on this method will involve implicit electrostatics at the boundary (which is missing in the current implementation) as well as more sophisticated treatments of the surface geometry (α shapes^{25,49} and tight Cocone).⁵⁰ The nonconvex hull geometries are significantly more expensive [$\mathcal{O}(N^2)$] than the convex hull [$\mathcal{O}(N \log N)$] but would enable the use of hull volumes directly in computing the compressibility of the sample.

APPENDIX A: COMPUTING CONVEX HULLS ON PARALLEL COMPUTERS

In order to use the Langevin Hull for simulations on parallel computers, one of the more difficult tasks is to compute the bounding surface, facets, and resistance tensors when the individual processors have incomplete information about the entire system's topology. Most parallel decomposition methods assign primary responsibility for the motion of an atomic site to a single processor, and we can exploit this to efficiently compute the convex hull for the entire system.

The basic idea involves splitting the point cloud into spatially overlapping subsets and computing the convex hulls for each of the subsets. The points on the convex hull of the entire system are all present on at least one of the subset hulls. The algorithm works as follows:

- (1) Each processor computes the convex hull for its own atomic sites (left panel in Figure 9).
- (2) The Hull vertices from each processor are communicated to all of the processors, and each processor assembles a complete list of hull sites (this is much smaller than the original number of points in the point cloud).
- (3) Each processor computes the global convex hull (right panel in Figure 9) using only those points that are the union of sites gathered from all of the subset hulls. Delaunay triangulation is then done to obtain the facets of the global hull.

The individual hull operations scale with $\mathcal{O}[(n/p) \log(n/p)]$, where n is the total number of sites, and p is the number of processors. These local hull operations create a set of p hulls, each with approximately $n/3(pr)$ sites for a cluster of operation r . The worst-case communication cost for using a "gather" operation to distribute this information to all processors is $\mathcal{O}[\alpha(p-1) + n\beta(p-1)/3rp^2]$, while the final computation of the system hull scales as $\mathcal{O}[(n/3r) \log(n/3r)]$.

For a large number of atoms on a moderately parallel machine, the total costs are dominated by the computations of the individual hulls, and communication of these hulls to create the Langevin Hull sees roughly linear speed-up with increasing processor counts.

AUTHOR INFORMATION

Corresponding Author

*E-mail: gezelter@nd.edu.

ACKNOWLEDGMENT

Support for this project was provided by the National Science Foundation under grant CHE-0848243. Computational time

was provided by the Center for Research Computing (CRC) at the University of Notre Dame.

Molecular graphics images were produced using the UCSF Chimera package from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIH P41 RR001081).

REFERENCES

- (1) Melchionna, S.; Ciccotti, G.; Holian, B. L. *Mol. Phys.* **1993**, *78*, 533–544.
- (2) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695.
- (3) Andersen, H. C. *J. Chem. Phys.* **1980**, *72*, 2384–2393.
- (4) Sturgeon, J.; Laird, B. J. *J. Chem. Phys.* **2000**, *112*, 3474–3482.
- (5) Berendsen, H. J. C.; Postma, J. P. M.; VanGunsteren, W. F.; Dinola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
- (6) Feller, S. E.; Zhang, Y. H.; Pastor, R. W.; Brooks, B. R. *J. Chem. Phys.* **1995**, *103*, 4613–4621.
- (7) Jakobsen, A. *J. Chem. Phys.* **2005**, *122*.
- (8) Asthagiri, D.; Paliwal, A.; Abras, D.; Lenhoff, A.; Paulaitis, M. *Biophys. J.* **2005**, *88*, 3300–3309.
- (9) Brown, G. C. *J. Theor. Biol.* **1991**, *153*, 195–203.
- (10) Berkowitz, M.; McCammon, J. A. *Chem. Phys. Lett.* **1982**, *90*, 215–217.
- (11) Brooks, C. L., III; Karplus, M. *J. Chem. Phys.* **1983**, *79*, 6312–6325.
- (12) Warshel, A. *Chem. Phys. Lett.* **1978**, *55*, 454–458.
- (13) Belch, A.; Berkowitz, M. *Chem. Phys. Lett.* **1985**, *113*, 278–282.
- (14) Lee, C.; McCammon, J.; Rossky, P. *J. Chem. Phys.* **1984**, *80*, 4448–4455.
- (15) King, G.; Warshel, A. *J. Chem. Phys.* **1989**, *91*, 3647–3661.
- (16) Beglov, D.; Roux, B. *J. Chem. Phys.* **1994**, *100*, 9050–9063.
- (17) Beglov, D.; Roux, B. *Biopolymers* **1995**, *35*, 171–178.
- (18) Li, Y.; Krilov, G.; Berne, B. *J. Phys. Chem. B* **2005**, *109*, 463–470.
- (19) Zhu, W.; Krilov, G. *J. Mol. Struct. (THEOCHEM)* **2008**, *864*, 31–41.
- (20) Kohanoff, J.; Caro, A.; Finnis, M. *ChemPhysChem* **2005**, *6*, 1848–1852.
- (21) Baltazar, S. E.; Romero, A. H.; Rodriguez-Lopez, J. L.; Terrones, H.; Martonak, R. *Comput. Mater. Sci.* **2006**, *37*, 526–536.
- (22) Delaunay, B. *Bull. Acad. Sci. USSR VII: Cl. Sci. Math. Nat.* **1934**, 793–800.
- (23) Lee, D. T.; Schachter, B. J. *Int. J. Parallel Program.* **1980**, *9*, 219–242.
- (24) Barber, C. B.; Dobkin, D. P.; Huhdanpaa, H. T. *ACM Trans. Math. Software* **1996**, *22*, 469–483.
- (25) Edelsbrunner, H.; Mucke, E. P. *ACM Trans. Graphics* **1994**, *13*, 43–72.
- (26) Schlick, T. *Molecular Modeling and Simulation: An Interdisciplinary Guide*; Springer-Verlag, Inc.: Secaucus, NJ, 2002.
- (27) García de la Torre, J.; Huertas, M. L.; Carrasco, B. *Biophys. J.* **2000**, *78*, 719–730.
- (28) García de la Torre, J. *Biophys. Chem.* **2001**, *94*, 265–274.
- (29) García de la Torre, J.; Carrasco, B. *Biopolymers* **2002**, *63*, 163–167.
- (30) Sun, X.; Lin, T.; Gezelter, J. D. *J. Chem. Phys.* **2008**, *128*, 234107.
- (31) Meineke, M. A.; Vardeman, C. F.; Lin, T.; Fennell, C. J.; Gezelter, J. D. *J. Comput. Chem.* **2005**, *26*, 252–271.
- (32) Gezelter, J. D.; Kuang, S.; Marr, J.; Stocker, K. M.; Li, C.; Vardeman, C. F.; Lin, T.; Fennell, C. J.; Sun, X.; Daily, K.; Zheng, Y.; Meineke, M. A. *OpenMD, an open source engine for molecular dynamics*; University of Notre Dame: Notre Dame, IN, 2010; <http://openmd.net>. Accessed February 17, 2011).
- (33) Barber, C. B.; Huhdanpaa, H. *Qhull*; University of Minnesota: Minneapolis, MN, 1996; <http://www.qhull.org>. Accessed February 17, 2011).

- (34) Berendsen, H.; Grigera, J.; Straatsma, T. J. *Phys. Chem.* **1987**, *91*, 6269–6271.
- (35) Qi, Y.; Çağın, T.; Kimura, Y.; Goddard, W. A., III *Phys. Rev. B* **1999**, *59*, 3527–3533.
- (36) Fennell, C. J.; Gezelter, J. D. *J. Chem. Phys.* **2006**, *124*, 234104(12).
- (37) Wolf, D.; Keblinski, P.; Phillpot, S. R.; Eggebrecht, J. J. *Chem. Phys.* **1999**, *110*, 8254–8282.
- (38) Dou, Y.; Zhigilei, L.; Winograd, N.; Garrison, B. J. *Phys. Chem. A* **2001**, *105*, 2748–2755.
- (39) Foiles, S. M.; Baskes, M. I.; Daw, M. S. *Phys. Rev. B* **1986**, *33*, 7983–7991.
- (40) Sutton, A. P.; Chen, J. *Phil. Mag. Lett.* **1990**, *61*, 139–146.
- (41) Kimura, Y.; Qi, Y.; Çağın, T.; Goddard, W. A., III The Quantum Sutton-Chen Many Body Potential for Properties of fcc Metals. *Technical Report 003*; Caltech ASCI: Pasadena, CA, 1998.
- (42) Collard, S.; McLellan, R. *Acta metall. mater.* **1991**, *39*, 3143–3151.
- (43) Glättli, A.; Daura, X.; van Gunsteren, W. J. *Chem. Phys.* **2002**, *116*, 9811–9828.
- (44) Motakabbir, K.; Berkowitz, M. J. *Phys. Chem.* **1990**, *94*, 8359–8362.
- (45) Pi, H. L.; Aragoes, J. L.; Vega, C.; Noya, E. G.; Abascal, J. L.; Gonzalez, M. A.; McBride, C. *Mol. Phys.* **2009**, *107*, 365–374.
- (46) Fine, R. A.; Millero, F. J. *J. Chem. Phys.* **1973**, *59*, 5529–5536.
- (47) Debenedetti, P. J. *Chem. Phys.* **1986**, *84*, 1778–1787.
- (48) Taylor, R. S.; Dang, L. X.; Garrett, B. C. *J. Phys. Chem.* **1996**, *100*, 11720–11725.
- (49) Edelsbrunner, H. *Discrete Comput. Geom.* **1995**, *13*, 415–440.
- (50) Dey, T.; Giesen, J.; Goswami, S.; Zhao, W. *Discrete Comput. Geom.* **2003**, *29*, 419–434.

Coupled Cluster in Condensed Phase. Part I: Static Quantum Chemical Calculations of Hydrogen Fluoride Clusters

Joachim Friedrich,^{*,†} Eva Perl,[‡] Martin Roatsch,[‡] Christian Spickermann,^{‡,§} and Barbara Kirchner[‡]

[†]Institute for Chemistry, Chemnitz University of Technology, Strasse der Nationen 62, 09111 Chemnitz, Germany

[‡]Wilhelm-Ostwald-Institut für Physikalische und Theoretische Chemie, Universität Leipzig, Linnéstrasse 2, D-04103 Leipzig, Germany

ABSTRACT: A multiscale approach with roots in electronic structure calculations relies on the good description of intermolecular forces. In this study we lay the foundations for a condensed phase treatment based on the electronic structure of hydrogen fluoride on a very high level of theory. This investigation comprises cluster calculations in a static quantum chemical approach employing density functional theory, second order Møller–Plesset perturbation theory (MP2) and the coupled cluster singles, doubles with perturbative triples method in combination with several basis sets as well as at the complete basis set limit. The clusters we considered are up to 12 monomer units large and consist of ring and chain structures. We find a good agreement of the intramolecular distance obtained from the MP2 approach and the largest basis set. The binding energy of the hydrogen fluoride dimer calculated from coupled cluster at the basis set limit agrees excellently with experiment, whereas the calculated frequencies at all levels agree reasonably well with different experimental values. Large cooperative effects are observed for the ring structures as compared to the chain clusters. The energy per monomer unit indicates the most stable structures to be the ring clusters.

1. INTRODUCTION

Cluster structure calculations of associated compounds in terms of quantum chemical first principles methods are an established concept for gaining detailed information about the intermolecular interactions between the constituting molecules and the identification of local ordering patterns possibly relevant for the condensed phase as well.^{1–3} Thus, such calculations can constitute a scale-transferring approach if something about the condensed phase should be learned. Typical objectives addressed by such an *ab initio* treatment of isolated cluster structures are the analysis of hydrogen bonding and cooperative effects on the basis of sophisticated electronic structure methods.

Small- to medium-sized hydrogen fluoride (HF) clusters constitute a particular suitable system for such studies due to the small size of the molecule and the generic character of the $\text{FH}\cdots\text{F}$ hydrogen bond.^{3–7} The methods employed in many of these studies are often based on density functional theory (DFT) approaches or second order Møller–Plesset perturbation (MP2) theory, whereas a more elaborate treatment of electron correlation is rarely applied for systems larger than the HF monomer or dimer.^{3,4,8–15}

An extensive study by Maerker et al. demonstrates that many structural quantities are well reproduced by DFT methods (particular hybrid functionals), but in general these approaches are not capable to obtain the accuracy of a post Hartree–Fock treatment in combination with large basis sets.³ A good agreement between structural results obtained from hybrid DFT calculations and experimental data was also reported by Guedes et al., but the interaction energies calculated for larger clusters were found to overestimate the experimental references.⁷ Additional examinations of cooperative effects in HF clusters have been carried out in terms of a simple relation between computed interaction energies and the cluster size. The largest cooperative effects were found for the cyclic trimer and tetramer structures.⁷

Similar results were also reported for MP2 computations on small- to medium-sized HF clusters, and it has been stated that such nonadditive contributions in the intermolecular interactions are important for a proper treatment of HF in the condensed phase as well.^{5,16}

In addition, isolated cluster calculations (and the corresponding energies) are frequently employed for the generation or modification of analytical force fields used in sampling methods, like traditional molecular dynamics simulations or Monte Carlo studies.^{16–26} In the case of liquid HF, the design of such force fields on the basis of *ab initio* calculations (or experimental data) proved to be difficult, which is not necessarily to be expected due to the putative simplicity of this molecule.^{16,20,23} Several experimental and first principles molecular dynamics studies indicate that there is an abundance of staggered linear chains in the liquid phase of HF and that the distance between these chains is relatively large.^{27,28} It has been stated that this local inhomogeneity in the intermolecular interactions (i.e., the difference between F–F intrachain and F–F interchain distances) largely contributes to the difficulties arising in the force field design.²⁰ The local bend structure found in the HF dimer (and also in the staggered chains) is known to be a result of the permanent multipole interactions between the hydrogen-bonded monomer units.²⁹ Such interactions are expected to be well captured in isolated cluster calculations and can be easily accounted for in analytical force fields. However, the weak interactions arising between HF molecules from different chains have been found to influence the density of the liquid phase to a considerable extent, which is the reason why these medium-range forces also have to be considered in the design of a force field and for the computational treatment of liquid HF in general. In the case of isolated

Received: March 10, 2010

Published: March 18, 2011

cluster studies such effects are only partly accounted for, but a more extended treatment of such intercluster effects is possible in the frame of the quantum cluster equilibrium (QCE) approach.^{30,32} The QCE method is in principle a multilevel or multiscale approach, because it combines static isolated molecule, i.e., quantum chemical, calculations with basic statistical mechanics in order to predict thermodynamics of the condensed phase. Therefore, it is possible to apply high-level electronic structure methods to the condensed phase.

The present study gives a detailed analysis of small- to medium-sized isolated cluster structures $(\text{HF})_n$ ($n \leq 12$) on the basis of high-quality coupled cluster methods, which have not been applied to HF clusters of this size before. The effect of the basis set size is investigated in terms of a complete basis set (CBS) extrapolation and compared to numbers obtained from DFT calculations.

The question whether ring structures play a significant role in liquid HF besides chain structures has been previously addressed in several experimental as well as theoretical studies.^{16,19,20,27,28,33–36} Whereas ab initio molecular dynamics simulations indicate a predominance of chain-like structures in liquid HF at ambient as well as supercritical states,^{16,35,36} experimental studies do not definitely attest this observation but rather assume an equality between the structure of the solid (parallel zigzag chains) and the structure of the liquid.³³ A recent first principles Monte Carlo study of the HF vapor phase indicates that for smaller aggregates (between three to six monomers) the cyclic arrangement is of increased importance and that the bulk vapor phase also contains a considerable amount of noncyclic structures.^{26,37} These findings are in contrast to earlier measurements in which the vapor phase was found to consist solely of cyclic clusters.^{38,39} In addition, experimental investigations of the liquid phase explicitly discuss the occurrence of cyclic species and highlight that cooperative effects, which are assumed to play only a minor role in the liquid as compared to the gas phase, could as well seriously affect the liquid phase structures of strongly associated liquids, such as water or HF.²⁷ In order to account for both structural motifs, clusters of chain- and ring-like geometry have been calculated for a subsequent QCE application.

The results of this study form the ab initio part of a multiscale approach toward the determination of (macroscopic) thermodynamic properties of liquid HF in terms of the QCE model over a large temperature domain, which will be presented in the second part of this study. In this way, a sophisticated post Hartree–Fock treatment of the condensed phase is possible, and the effect of electron correlation on thermodynamic properties of the condensed phase can be investigated in detail.

The study is organized as follows. First, the computational details of the cluster calculations are summarized with special regard to the coupled cluster calculations and the incremental scheme applied for the larger cluster structures. The following section introduces the set of investigated clusters and summarizes the results obtained for geometries and energies. A comparison between calculated harmonic frequencies and anharmonic as well as experimental frequencies is presented next. The paper closes with a discussion of the results and a conclusion.

2. THEORY

2.1. Computational Details. In order to evaluate molecular properties of the different clusters their geometries, harmonic frequencies, and interaction energies have to be computed. The

intracluster interaction energies $E_{\mathcal{P}}^{\text{intra}}$ are obtained according to the supramolecular approach:

$$E_{\mathcal{P}}^{\text{intra}} = E_{\mathcal{P}} - i(\mathcal{P})E_1 \quad (1)$$

where $E_{\mathcal{P}}$ and E_1 denote the total energies of the clusters containing $i_{\mathcal{P}}$ monomers and the corresponding monomer in its relaxed geometry, respectively (adiabatic interaction energy). For the determination of cluster interaction energies according to eq 1, structure optimizations were performed employing DFT as well as MP2 theory with the resolution of identity (RI) procedure.⁴⁰ Interaction energies have not been corrected for the zero-point energy. The program packages used for the actual ab initio calculations were TURBOMOLE 5.91 and associated programs.⁴¹ For DFT calculations, the gradient-corrected functionals B-P86 and PBE were employed in combination with the TZVP basis set as well as the RI technique, whereas MP2 calculations were carried out for the TZVP and QZVP basis sets.^{42–48} In order to obtain accurate structures for the CCSD(T) single point calculations (see Section 2.2), the convergence criterion for the MP2/QZVP optimization was set to 10^{-5} atomic units for the norm of the Cartesian gradient. These data will be abbreviated as MP2/QZVP* or as the basis set QZVP* only. The retrieved cluster interaction energies were counterpoise (CP) corrected by the generalization of the Boys and Bernardi scheme as introduced by Wells and Wilson.^{49,50} The other binding energies being relevant for the QCE calculations are the binding energy per hydrogen bond:

$$E_{\mathcal{P},\text{hbond}}^{\text{intra}} = E_{\mathcal{P}}^{\text{intra}} / n_{\text{hbond}} \quad (2)$$

with n_{hbond} being the number of hydrogen bonds in the cluster.

For the determination of the harmonic frequencies, the SNF program package was employed after the electronic structure calculations were carried out.⁵¹ The SNF program computes frequencies on the basis of the harmonic approximation as numerical derivatives of the analytic gradients provided by the structure optimization. All harmonic frequencies entered the vibrational partition function unscaled. We neglected anharmonic effects. Such effects, for example, nuclear quantum effects for the proton transfer or isotope effects, have been broadly studied by the group of Hammes-Schiffer.^{52–54}

2.2. Computational Details of Coupled Cluster Calculations. The coupled cluster singles and doubles with perturbative triples CCSD(T) energies were calculated with the MOLPRO quantum chemistry package^{55–57} using the aug-cc-pVXZ ($X = \text{T}, \text{Q}, \text{5}$) basis sets.^{58,59} For the larger clusters the incremental scheme⁶⁰ had to be applied in order to keep the calculations in the aug-cc-pVQZ basis feasible (756, 1008, 1512 basis functions in C_1 symmetry). Here the total correlation energy is obtained from correlation calculations in small domains. In order to account for the nonadditive effects, one has to calculate correlation contributions from pairs, triples, and so on, of domains until the desired accuracy is reached. The total correlation energy is then obtained by the incremental series:

$$E_{\text{corr}} = \sum_i \Delta \varepsilon_i + \frac{1}{2!} \sum_{ij} \Delta \varepsilon_{ij} + \frac{1}{3!} \sum_{ijk} \Delta \varepsilon_{ijk} + \dots \quad (3)$$

$$\Delta \varepsilon_i = \varepsilon_i, \quad \Delta \varepsilon_{ij} = \varepsilon_{ij} - \Delta \varepsilon_i - \Delta \varepsilon_j$$

Table 1. CCSD(T) and MP2 Adiabatic Interaction Energies $E_{\mathcal{P}}^{\text{intra}}$ for the HF Dimer Using Different Basis Sets of the aug-cc-pVXZ Series^a

basis	SCF		MP2		CCSD(T)	
	$E_{\mathcal{P}}^{\text{intra}}$	CP	$E_{\mathcal{P}}^{\text{intra}}$	CP	$E_{\mathcal{P}}^{\text{intra}}$	CP
TZVP	-17.52	1.02	-20.10	3.11		
QZVP*	-14.85	0.30	-19.65	1.84		
aug-cc-pVDZ	-15.44	1.05	-19.58	2.96	-20.15	3.40
aug-cc-pVTZ	-14.81	0.50	-19.69	2.06	-20.21	2.18
aug-cc-pVQZ	-14.76	0.23	-19.38	1.10	-19.77	1.00
aug-cc-pVSZ	-14.57	0.03	-19.15	0.68	-19.46	0.53
aug-cc-pV6Z	-14.54	0.00	-18.95	0.40	-19.27	0.27
cc-pVDZ-F12 ^b	-14.64	0.18	-18.55	1.09	-18.71	1.33
cc-pVTZ-F12 ^b	-14.63	0.10	-18.80	0.40	-19.18	0.55
cc-pVQZ-F12 ^b	-14.57	0.02	-18.77	0.14	-19.16	0.19
	extrapolation					
CBS(23)			-19.99		-20.50	
CBS(34)			-19.19		-19.48	
CBS(45)			-19.12		-19.35	
CBS(56)			-18.71		-19.05	

^aCP denotes counterpoise correction. All energies in kJ/mol.

^bSCF+CABS singles, MP2-F12, and CCSD(T)(F12) energies.

where ε_i is the correlation energy of the i -th domain and $\Delta\varepsilon_{ij}$ is the two-body correction to the correlation energy for the domains i and j together.

The calculations were performed with the fully automated implementation of the incremental scheme^{61,62} using the domain-specific basis set approach.^{63,64} The incremental expansion was truncated at third order, the domain size parameter was set to 4 orbitals, and the connectivity parameter was set to 3 Bohr in order to obtain whole HF molecules as one-site domains. The order-dependent energy convergence threshold⁶⁵ for the accuracy of the coupled cluster calculations in the domains was set to $10^{-6} E_h$. The parameter to determine the environment of a domain (t_{main})^{63,64} was set to 3 Bohr. In the environment of a domain the basis set of the hydrogens was reduced to STO-3G and to 6-31G⁶⁶ for fluorine. In all calculations the frozen core approximation was applied to the 1s orbitals of fluorine. For R8a and R₂6 (see Figure 2) an order-dependent distance screening with $t_{\text{dist}} = 30/(\mathcal{O}_i - 1)^2$ Bohr was applied for the two- and three-body increments, where \mathcal{O}_i is the order of the increment.^{63,62}

The extrapolation to the basis set limit was performed with the two point scheme of Halkier and Helgaker et al.⁶⁷ For the small clusters we performed additional explicitly correlated MP2 and CCSD(T) calculations with the TURBOMOLE 6.2 package^{68,69} using the cc-pVQZ-F12 basis set of Peterson et al.⁷⁰ with the corresponding CABS⁷¹ and the aug-cc-pwCVSZ basis set for density fitting.⁷² Furthermore the Ansatz 2 was employed together with the approximation B and the fixed amplitudes formulation.⁷³ The exponent of the correlation factor⁷⁴ that determines the F12 basis was set to the recommended value of $1.1a_0^{-1}$. The Hartree–Fock energies were corrected with the CABS singles.⁷⁵

3. RESULTS

3.1. Benchmarking the Accuracy. The CCSD(T) adiabatic interaction energies have to be calculated using large basis sets,

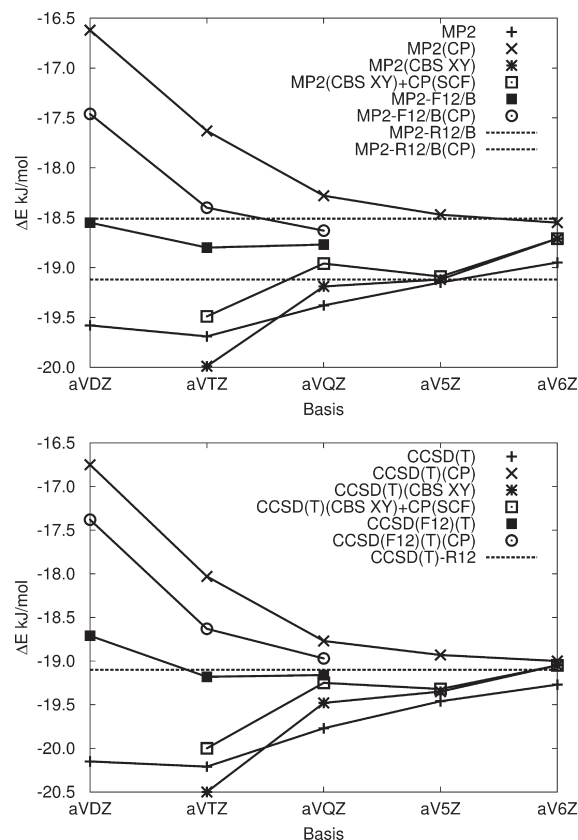


Figure 1. Convergence of the adiabatic interaction energies $E_{\mathcal{P}}^{\text{intra}}$ for MP2 and CCSD(T) with respect to the basis set. The explicitly correlated MP2-R12 and CCSD(T)-R12 values were taken from ref 4. The CCSD(T)(F12) and MP2-F12 energies were calculated with the cc-pVXZ-F12 basis sets using the recommended γ of 0.9, 1.0, and 1.1 a_0^{-1} , respectively. Additionally the CABS singles correction has been included.

including diffuse functions, to obtain the required accuracy. On the other hand the calculations still have to be feasible for the large clusters, which limits the basis set to aug-cc-pVQZ. In order to improve the accuracy of the coupled cluster energies, we use the 34 extrapolation to the basis set limit. This procedure was recently used to benchmark interaction energies for hydrogen bonds in DNA base pairs.⁷⁶ However, to validate this procedure for HF clusters, we investigate the HF dimer with the aug-cc-pVXZ basis set series ($X = D, T, Q, 5, 6$) in combination with two point extrapolations^{67,77,78} from 23 to 56 (Table 1). Such extrapolations using extended basis sets are frequently used to benchmark the explicitly correlated F12 methods.^{68,79} To obtain a further theoretical reference, we calculated the MP2-F12 and CCSD(T)(F12) adiabatic interaction energies for the cc-pVXZ-F12⁷⁰ ($X = D, T, Q$) basis set series including counterpoise correction. Since the HF dimer was intensively studied in the past,^{3,4,80–82,83} we can furthermore use the adiabatic interaction energies of the dimer to validate the accuracy of the 34 extrapolation.

The convergence of the adiabatic interaction energies with respect to the basis set is given in Figure 1. The non-CP-corrected curve converges from below to the CBS limit and the CP-corrected curve converges from above for MP2 and for CCSD(T), as expected from theory. Furthermore the two-point extrapolated adiabatic interaction energies are in between the two bounds

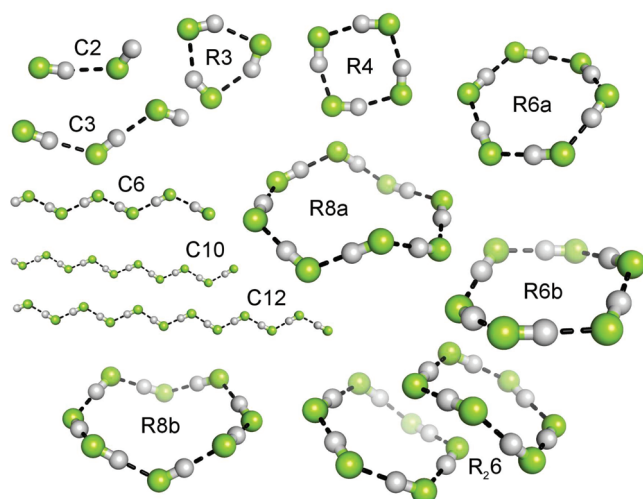


Figure 2. HF clusters investigated in the present study. Depicted are the MP2/TZVP optimized structures. First character R denotes a ring structure and first character C a chain structure. Please note that there is a boat-like (R6a) and a chair-like (R6b) conformation of the hexamer ring cluster.

except for the 23 extrapolated MP2 and CCSD(T) energies. Comparing the non-CP-corrected interaction energies in Table 1 with the R12 results of Klopper et al.⁴ (MP2-R12/A, -18.89 kJ/mol; MP2-R12/B, -19.12 kJ/mol; CCSD(T)-R12, -19.26 kJ/mol), we find a good agreement with our calculations using the quintuple- or hexuple- ζ basis sets. Also the estimate for the complete basis set limit of -19.1 kJ/mol for CCSD(T) agrees very well with our 56-extrapolated value of -19.05 kJ/mol. Furthermore this value compares very well to the value extracted from the experimental D_0 of Miller⁸⁴ (-19.1 kJ/mol) and is consistent with the previous work of Pine and Howard (-19.08 kJ/mol).⁸⁵ Considering the calculations of Boese et al.,⁸⁶ we find a good agreement to the nonrelativistic frozen core interaction energies of -19.11 kJ/mol. However, the interaction energy changes by a few tenths of a kJ/mol (0.29 , -0.33 kJ/mol) by correlating all electrons and by including relativistic effects.⁸⁶ Since these effects have the same order of magnitude but different signs it is evident that the good agreement of the nonrelativistic 56-extrapolated frozen core CCSD(T) energy with the experimental result is due to a cancellation of both effects. On the other hand both effects are small, and therefore they are neglected. Comparing the 34-extrapolated energies with the best estimate of this work we find a difference of only 0.43 kJ/mol, which is an error on the same order of the frozen core approximation or relativistic effects. This error is smaller than the error of the applied DFT methods with -1.11 and 3.35 kJ/mol for B-P86 and PBE, respectively. Please note that the results of the explicitly correlated MP2 and CCSD(T) using the cc-pVQZ-F12 basis set agree very well to our CBS(56). Since the CP correction is small (0.14 kJ/mol for MP2, 0.19 kJ/mol for CCSD(T)) this holds for the CP-corrected and the uncorrected adiabatic interaction energy. To obtain higher accuracy one has to include other effects, like core correlation or relativistic effects as well, which is beyond the scope of this work. Based on the considerations above, we conclude that the 34-extrapolated energies are sufficiently accurate for our purposes.

3.2. Systems Investigated. The present cluster sizes (HF)_{*n*} vary from $n = 1$ for the monomer to $n = 12$ for the dimer of the cyclic hexamer as the largest cluster similar as in previous studies.^{3,21,35,87}

Table 2. Average Interatomic Distances r and Angles α of All Investigated Clusters for the MP2 Electronic Structure Methods^a

cluster	MP2							
	TZVP				QZVP			
	$r(\text{HF})$	$r(\text{H}\cdots\text{F})$	$r(\text{FF})$	$\alpha(\text{FHF})$	$r(\text{HF})$	$r(\text{H}\cdots\text{F})$	$r(\text{FF})$	$\alpha(\text{FHF})$
C2	92.7	184.8	276.5	169.8	92.2	181.2	272.6	170.7
C3	92.9	176.4	269.3	175.7	92.5	173.8	266.4	175.8
C6	93.7	165.9	259.8	179.3	93.2	164.5	258.0	179.0
C10	94.3	160.6	255.0	179.0	—	—	—	—
C12	94.5	159.2	253.8	178.9	—	—	—	—
R3	93.6	181.5	262.9	143.9	93.4	175.6	259.1	147.3
R4	94.8	161.1	252.9	161.7	94.5	158.8	251.2	164.7
R6a	95.4	153.4	248.6	175.7	94.9	152.8	247.6	177.6
R6b	95.4	153.4	248.6	175.7	95.0	152.7	247.6	176.7
R8a	95.5	152.2	247.6	176.4	95.0	152.2	247.1	177.8
R8b	95.5	152.3	247.8	176.5	95.0	152.1	247.1	177.6
R _{2,6}	95.5	153.7	248.9	174.9	95.5	153.6	248.8	174.9

^a All distances in pm and angles in degrees.

For a complete overview of the investigated clusters cf. Figure 2. In order to characterize the stationary points illustrated in Figure 2, the eigenvalues of the Hessian were examined. From these it is found that all structures are true minima on the corresponding potential energy surfaces (PES) with the exception of R6a (transition state per B-P86/TZVP) and C6 (third-order saddle point per MP2/QZVP). The C10 cluster is a minimum only for the MP2/TZVP method, and C12 is found to be a minimum in case of the MP2/TZVP and B-P86/TZVP calculations. It should be noted that in contrast to the finite-temperature first principles Monte Carlo sampling of ref 37, no branched cluster structures were investigated in the zero Kelvin cluster optimizations of the present study.

3.3. Geometries. The geometric parameters (hydrogen bond) for all examined clusters are given in Tables 2–4.

The $r(\text{HF})$ distance of the monomer amounts 91.627 pm for MP2/QZVP, 91.627 pm for MP2/QZVP*, 92.159 pm for MP2/TZVP, 93.270 pm for B-P86/TZVP, and 93.246 pm for PBE/TZVP. Obviously both MP2/QZVP results compare best to the experimental value 91.680 pm given in ref 20. There are some values for the HF monomer equilibrium distance based on high-level quantum chemical calculations in the literature, yielding values of 91.69 ,⁸⁸ 91.70 ,⁸⁹ and 91.708 pm.⁹⁰ The authors of the first two studies point out that in the case of HF, the good performance of the CCSD(T) approach in combination with medium-sized basis sets stems from an error cancellation between basis set truncation and truncation of the excitation level in the applied coupled cluster methodology. A similar situation is found for the MP2 equilibrium distances listed in ref 89. Whereas many values obtained from the conventional MP2 calculation in combination with medium-sized basis sets are within ~ 0.1 pm of the experimental result, somewhat larger discrepancies are found for the MP2-R12 approach, i.e., for taking a substantial step toward basis set saturation. Thus, the best estimate of 91.627 pm (MP2/QZVP) from our calculations could benefit from a similar error cancellation as well and thereby approach the experimental reference as close as 0.06 pm.

Table 3. Average Interatomic Distances r and Angles a of All Investigated Clusters for MP2/QZVP*^a

cluster	MP2/QZVP*			
	$r(\text{HF})$	$r(\text{H}\cdots\text{F})$	$r(\text{FF})$	$a(\text{FHF})$
C2	92.1	181.2	272.6	170.7
C3	92.5	173.9	266.4	175.6
C6	93.2	164.5	258.0	179.0
R3	93.4	175.6	259.1	147.3
R4	94.5	158.8	251.2	164.7
R6a	94.9	152.7	247.6	177.2
R6b	95.1	152.7	247.5	176.7
R8a	95.1	152.1	247.1	177.7
R8b	95.0	152.1	247.0	177.7
R ₂ 6	95.0	152.9	247.7	176.8

^a All distances in pm and angles in degrees.**Table 4.** Average Interatomic Distances r and Angles a of All Investigated Clusters for the DFT Methods^a

cluster	DFT							
	B-P86/TZVP				PBE/TZVP			
	$r(\text{HF})$	$r(\text{H}\cdots\text{F})$	$r(\text{FF})$	$a(\text{FHF})$	$r(\text{HF})$	$r(\text{H}\cdots\text{F})$	$r(\text{FF})$	$a(\text{FHF})$
C2	93.9	180.0	272.3	165.7	93.9	180.0	271.6	163.5
C3	94.5	171.2	267.2	173.8	94.4	171.2	265.5	173.4
C6	95.8	159.4	255.5	178.3	95.7	159.5	255.6	178.3
C12	97.4	150.6	248.3	179.3	—	—	—	—
R3	96.4	167.1	253.9	147.8	96.2	168.3	254.7	147.5
R4	98.9	147.8	244.7	165.1	98.6	148.7	245.3	165.0
R6a	100.2	140.5	240.6	177.8	100.0	140.8	240.7	176.7
R6b	100.2	140.4	240.5	176.8	100.1	140.7	240.6	176.3
R8a	100.4	139.5	239.9	177.8	100.2	140.0	240.1	177.3
R8b	100.4	139.4	239.8	177.8	100.3	139.6	239.8	177.5
R ₂ 6	100.2	140.4	240.5	176.8	100.0	141.2	241.1	177.1

^a All distances in pm and angles in degrees.

For all methods we observe an increase in the intramolecular HF distance the larger the clusters are, which is also found in finite-temperature first principles calculations.³⁷ Furthermore, this bond elongation is stronger for the cyclic structures than for the linear geometries. The hydrogen-bond distances become smaller the larger the clusters are, and they are smaller in cyclic structures than in linear geometries. Both of these observations point to the fact that the hydrogen bonds are stronger with increasing number of monomers in the clusters and in comparing cyclic with linear structures. Chaban and Gerber calculated for the cyclic trimer R3 values of $r(\text{HF}) = 93.0$ pm and $r(\text{H}\cdots\text{F}) = 185.7$ pm using MP2 in combination with a TZP basis set.⁹¹ For the cyclic tetramer the authors obtained $r(\text{HF}) = 94.0$ pm and $r(\text{H}\cdots\text{F}) = 165.9$ pm using the same method and basis set combination, thus these authors find the same trend as apparent in our data.⁹¹

Comparing the different methodologies, we find shorter distances with increasing the basis set for MP2 and a more linear arrangement of the hydrogen-bond angle indicating stronger hydrogen bonds, see Table 2. The changes with the tighter convergence criterion are marginal on average, see Table 3. The

hydrogen-bond distances are even shorter (B-P86 < PBE), and the intramolecular $r(\text{HF})$ distance is shorter (B-P86 > PBE) for DFT than for MP2. However, the MP2 data provide more linear hydrogen bonds; compare the $a(\text{FHF})$ angles in Tables 2 and 3 with those in Table 4.

3.4. Energies. The adiabatic interaction energies were computed by DFT as well as the MP2 method according to eq 1. Please note that the obtained geometries vary with regard to the applied quantum chemical methodology, for a discussion see Section 2.1. As an exception the CCSD(T)/CBS(34) and MP2/CBS(34) values are obtained at the QZVP* geometries. The results of the adiabatic interaction energies $E_{\text{int}}^{\text{intra}}$ are listed in Table 5. In order to validate the accuracy of the CBS(34) energies, we compare these results with explicitly correlated methods at MP2 and CCSD(T) levels. At the MP2 level the largest deviation between the explicitly correlated result and the CBS(34) is 2.53 kJ/mol for the tetramer. At the CCSD(T) level the largest deviation is 2.17 kJ/mol. As a further benchmark we calculated the CBS(45) for the C2, C3, and R3 clusters. The CCSD(T)/CBS(45) adiabatic interaction energies agree within 0.65 kJ/mol or better with the results from the explicitly correlated CCSD(T). Due to this good agreement of the two benchmarks, we conclude that the deviation of the CBS(34) to the CBS(45) is similar to the deviations from explicit correlation (vide supra).

Comparing CCSD(T)/CBS(34) with the CP-corrected results from Klopper et al.⁸² [$(\text{HF})_2 = 19.0$, $(\text{HF})_3 = 62.8$, and $(\text{HF})_4 = 113.5$] we find a maximum deviation of 4.9 kJ/mol for the tetramer. Comparing to the best estimate of Klopper et al.,⁸² the deviation drops down to 2.4 kJ/mol and is within the provided error bars of 3 kJ/mol. Comparing Kloppers values to our CCSD(T)(F12)/cc-pVQZ-F12 results, we find the largest deviation to be 3.3 kJ/mol with respect to the CP corrected energies and a deviation of 0.8 kJ/mol with respect to the best estimate (tetramer). Please note that the geometries of Klopper et al. are slightly different from those of this work. Therefore this comparison of the adiabatic interaction energies cannot be rigorously made. On the other hand the agreement is good and demonstrates the applicability of our computational approach.

Comparing the data for B-P86 functional and the MP2 results, a reasonable agreement is found in case of the smaller clusters, whereas the values obtained from the PBE functional are significantly larger in magnitude. The coupled cluster values are in between the MP2 and the DFT values. Concerning the larger cluster structures, a strong overbinding of the DFT methods as compared to the CCSD(T) and MP2 results is observed, which was already discussed by Maerker et al.³ Although all numbers are BSSE-corrected according to the counterpoise method, a relatively strong basis set dependence for the MP2 calculations is still apparent, which amounts to a difference of up to 44 kJ/mol for the R8a cluster. It should be noted here that a better adjusted basis set for MP2 was not considered for the sake of computational applicability, because all cluster structures are optimized in the chosen basis set. Comparing our DFT values to the ones previously published, we generally find a satisfactory agreement in the range of ± 10 kJ/mol in most cases.^{5,7,21} This is also true for the estimated dimer and trimer aggregation energies from the DFT Monte Carlo calculations of the HF vapor phase at finite temperatures.³⁷ Remaining differences can possibly be attributed to the smaller double- ζ basis sets employed for the geometry optimizations in some of the previous calculations.^{7,5}

Table 5. Adiabatic Interaction Energies $E_{\mathcal{L}}^{\text{intra}}$ for All Investigated Clusters and Different Electronic Structure Methods^a

cluster	$i(\mathcal{L})$	CCSD(T)			MP2			B-P86	PBE		
		cc-pVQZ-F12	CBS(45)	CBS(34)	cc-pVQZ-F12	CBS(34)	QZVP*	QZVP	TZVP	TZVP	TZVP
C2	2	-19.16	-19.12	-19.48	-18.77	-19.19	-17.81	-17.81	-16.99	-17.94	-21.40
C3	3	-44.32	-44.32	-45.13	-43.58	-44.60	-41.50	-41.49	-39.31	-43.12	-50.13
R3	3	-63.94	-63.29	-65.12	-62.68	-63.99	-59.15	-59.14	-50.52	-64.40	-72.57
R4	4	-116.84	–	-119.01	-115.90	-118.43	-109.77	-109.77	-94.98	-124.50	-136.49
C6	6	–	–	–	–	–	–	–	-118.32	-137.85	-154.94
R6a	6	–	–	-198.72	–	-198.85	-184.29	-184.28	-165.10	-212.36	-231.69
R6b	6	–	–	-199.05	–	-199.19	-184.59	-184.58	-165.19	-212.92	-232.40
R8a	8	–	–	-268.81	–	-269.00	-249.43	-248.98	-225.06	-288.93	-314.42
R8b	8	–	–	-269.11	–	-269.40	-250.06	-249.95	-225.55	-290.31	-316.41
C10	10	–	–	–	–	–	–	–	-231.00	–	–
C12	12	–	–	–	–	–	–	–	-288.31	-351.06	–
R ₂ 6	12	–	–	-415.07	–	-412.41	-377.83	-377.84	-331.07	-421.55	-465.95

^a All energies in kJ/mol. Some results are missing because the corresponding geometries are not minimum structures or the computation was infeasible on our hardware (CCSD(F12)(T), CCSD(T)/aug-cc-pV5Z).

Table 6. Adiabatic Interaction Energies $E_{\mathcal{L}, \text{hbond}}^{\text{intra}}$ per Hydrogen Bond for All Investigated Clusters and Different Electronic Structure Methods^a

cluster	$i(\mathcal{L})$	CCSD(T)		MP2			B-P86	PBE
		CBS(34)	CBS(34)	QZVP*	QZVP	TZVP	TZVP	TZVP
C2	2	-19.48	-19.19	-17.81	-17.81	-16.99	-17.94	-21.40
C3	3	-22.57	-22.30	-20.75	-20.75	-19.66	-21.58	-25.07
R3	3	-21.71	-21.33	-19.72	-19.71	-16.84	-21.47	-24.19
R4	4	-29.75	-29.61	-27.44	-27.44	-23.75	-31.13	-34.12
C6	6	–	–	–	–	-23.66	-27.57	-30.99
R6a	6	-33.12	-33.14	-30.72	-30.71	-27.52	-35.38	-38.62
R6b	6	-33.18	-33.20	-30.77	-30.76	-27.53	-35.49	-38.73
R8a	8	-33.60	-33.63	-31.18	-31.12	-28.13	-36.12	-39.30
R8b	8	-33.64	-33.68	-31.26	-31.24	-28.19	-36.29	-39.55
C10	10	–	–	–	–	-25.67	–	–
C12	12	–	–	–	–	-26.21	-31.91	–
R ₂ 6	12	-34.59	-34.37	-31.49	-31.49	-27.59	-35.13	-38.83

^a All energies in kJ/mol. Some results are missing because the corresponding geometries are not minimum structures.

The examination of interaction energies per hydrogen bond (see Table 6) gives a comparison of the relative stability of the different cluster structures and sizes. There are several approaches for measuring the degree of cooperativity in cluster structures, and some of them have been applied to HF clusters previously. Rincón et al. correlated cooperative effects obtained from bond distance and interaction energies to the critical point of the hydrogen bond in the topology of the electron density and found a linear relationship between these measures.⁵ By analyzing the interaction energy per hydrogen bond of cyclic HF clusters as well as the reaction energy for adding one HF monomer to a cluster of given size, these authors report considerable cooperative effects for the ring sizes $n = 3$ and 4. Similar observations can be found in a cluster study by Guedes et al. in which larger clusters of up to $n = 10$ monomer units are considered.⁷ Both studies also indicate a saturation of cooperative effects for cyclic HF clusters of size $n = 6$ and larger. If one considers the interaction energy per hydrogen bond (see

Table 6) as a measure for cooperativity, then it is immediately apparent that all clusters larger than the dimer are stabilized by cooperative effects. The largest stabilization can be found in case of the large cyclic clusters R6a/b and R8a/b for all investigated methods. If, however, one is interested in the variation of cooperativity with cluster size (i.e., in the difference in interaction energy per hydrogen bond for a given pair of n and $n - 1$), then it is seen from Table 6 that the largest change in hydrogen bond stabilization of ≈ 7 – 10 kJ/mol occurs for the transition R3 \rightarrow R4. Comparing the cooperativity between chain and ring structures for a given cluster size, it is apparent that the stabilization of the cyclic compounds is larger than those of the chain clusters for larger n , whereas the situation is reversed in the case of the trimer structures. This can be understood by considering a possible destabilization of the small trimer ring R3 in terms of ring strain.^{5,7}

As already discussed in the Introduction, the main motif of liquid HF predicted by ab initio simulations as well as recent

Table 7. Calculated Harmonic Frequencies from Different Quantum Chemical Methods, Comparison to Experimental Values, and Frequencies from Theoretical Investigations

cluster	MP2	B-P86	PBE	B3LYP	lit.	
	QZVP*	TZVP	TZVP	TZVP	aug-cc-pVDZ	
C1	4185	4159	3964 3956 ⁸⁹	3970	4073 ⁸⁹	expt 4138 ^{89,95} CC 4141, ⁹⁶ 4139, ⁸⁸ 4137, ⁸⁹ 4152 ⁸¹ MP2-R12 4138 ⁸⁹
C2	169	157	166	166	167 ⁹⁷	MP2-R12 155 ⁸⁹
C2	236	219	221	218	233 ⁹⁷	MP2-R12 209 ⁸⁹
C2	486	455	469	463	465 ⁹⁷	MP2-R12 467 ⁸⁹
C2	588	552	598	600	578 ⁹⁷	MP2-R12 547 ⁸⁹
C2	4049	4050	3772	3780	3882 ⁹⁷	MP2-R12 4030 ⁸⁹
C2	4140	4122	3921	3924	4016 ⁹⁷	MP2-R12 4090 ⁸⁹

experiments are linear zigzag chains.^{16,28,35} In the case of the small- to medium-sized clusters investigated in the present study, it is apparent that this predominance of chain structures is not reflected in the calculated interaction energies (see Table 5) or the cooperative effects (see Table 6). However, earlier studies by Karpfen et al. have shown that on lower theoretical levels the hydrogen-bond stabilization in cyclic structures approaches that of an infinite HF zigzag chain with increasing ring size.^{92–94}

Thus, the energetic preference of ring clusters as seen in Tables 5 and 6 is apparently rooted in the finite chain length considered here and in the fact that rings have an additional hydrogen bond.

Two additional aspects can be expected to contribute to the stability of chain structures in the case of the liquid phase of HF. First, open chain arrangements will be entropically favored at finite temperatures over closed cyclic structures, which in turn are stabilized by an additional hydrogen bond in isolated molecule calculations. However, in the real liquid there will be no completely isolated chain endings, and medium-range interactions from neighboring chains are expected to stabilize the loose tails, thereby partly accounting for the enthalpic penalty which these structures are subject to in the isolated quantum chemical calculation. The consideration of these effects via a finite temperature and a meanfield intercluster interaction will be covered in the second part of this study in terms of the QCE model.

3.5. The Accuracy of the Spectra. Besides the cluster interaction energies discussed in Section 3.3, the accuracy of the vibrational frequencies for the examined cluster structures have been found to be of crucial importance for a successful application in the frame of the QCE model.^{30,31}

Table 7 presents the vibrational frequencies of HF and the HF-dimer at MP2, DFT, and CC levels.^{81,88,89,96} The quantitative calculation of the harmonic wavenumber of HF was the objective of several investigations in the past,^{81,88,89,96} and it was demonstrated that the amazingly high accuracy of 1 cm⁻¹ can be obtained if relativistic effects, core and core–valence correlation effects, and higher excitations in coupled cluster as well as large basis sets are used in the calculations.^{88,96} Since these effects may have different signs, they might cancel to a large extent. This is the reason why the MP2-R12 result from ref 89 (4138 cm⁻¹) agrees perfectly with the experimental value from ref 95 (4138 cm⁻¹). Since the MP2-R12 calculation is close to the MP2 basis set limit it is a good reference for the quality of the

basis set in our MP2 calculations. Comparing the harmonic MP2/TZVP wavenumber, we find a difference of 21 cm⁻¹. The increase of the basis set from TZVP to QZVP shifts the MP2 harmonic wavenumber further away from the basis set limit (error 47 cm⁻¹). This effect is not surprising since a similar behavior was observed at the CC-level before.^{98,99} For the MP2/QZVP harmonic wavenumbers of the dimer we find similar errors as compared to those in ref 80, which were obtained from an empirical refinement of the MP2-R12 potential energy surface (14–50 cm⁻¹).

4. CONCLUSION

We presented an extensive electronic structure study of hydrogen fluoride (HF) clusters containing 1–12 monomers, applying density functional theory and second-order Møller–Plesset perturbation theory as well as the coupled cluster method. As basis sets we chose the Ahlrichs basis sets for the optimization (TZVP and QZVP), and for the coupled cluster calculations we applied Dunning basis sets and extrapolated them to the complete basis set (CBS) limit. For all methods we optimized the structures except for the MP2/CBS(34) and for the coupled cluster results at the complete basis set limit. These calculations are based on MP2 structures with the combination of a tight optimization criterion denoted as MP2/QZVP*.

For the geometry of the monomer we found *r*(HF) distances of 91.627 pm for MP2/QZVP, 91.627 pm for MP2/QZVP* (higher convergence criterion), 92.159 pm for MP2/TZVP, 93.270 pm for B-P86/TZVP, and 93.246 pm for PBE/TZVP. Thus the MP2/QZVP structures compare best to the experimental value of 91.680 pm as well as to previous computations (91.69 pm) based on high-level coupled cluster approaches and large basis sets.^{20,88,89} For all methods the intramolecular HF distance was observed to become larger with increasing cluster size. Furthermore, this bond elongation was stronger for the cyclic structures than for the chain geometries. The hydrogen bonds became shorter with increasing cluster size, and they were shorter in cyclic structures than in the linear geometries. Both of these observations indicate stronger hydrogen bonds with increasing number of monomers in the clusters and comparing cyclic with linear structures. Stronger hydrogen bonding of clusters also manifested in shorter distances for MP2 with increasing basis set and a more linear arrangement of the hydrogen-bond angle as compared to the DFT results.

For the energies we observed excellent agreement of the 34 extrapolated complete basis set results with both the coupled cluster (−19.48 kJ/mol) as well as with the MP2 (−19.19 kJ/mol) calculations to the experimental binding energy of the dimer. In the case of the MP2 method the well-known large basis set dependency was observed. For example, the R8a cluster values obtained with TZVP (−225.06 kJ/mol) and QZVP (−248.98 kJ/mol) deviate by 24 kJ/mol from each other. For obvious reasons we found that in general ring structures are energetically preferred over chain structures in isolated molecule calculations for all methods and cluster sizes. Similar conclusions can be drawn for the frequencies. The best agreement is found for the MP2/QZVP* calculations. Comparing ring structures with chain structures, we find that cooperativity is more important in the ring structures than in chain clusters from the indicator of interaction energy.

As mentioned in the Introduction, we want to apply the highly accurate electronic structure results for our small- to medium-sized cluster set obtained in this article to determine thermodynamic properties of the condensed phase over a large temperature interval. The first step in a multiscale condensed phase description based on the resolution of the electronic structure consists of the accurate treatment of the underlying electronic structure problem, which has been accomplished in the present study. In the subsequent article we will show that this is a necessary prerequisite in order to obtain accurate and consistent results for the condensed phase of HF.

AUTHOR INFORMATION

Corresponding Author

*E-mail: bkirchner@uni-leipzig.de.

Present Addresses

[§]Lehrstuhl für Anorganische Chemie 2, Organometallics and Materials Chemistry, Ruhr-Universität Bochum, Universitätsstrasse 150, D-44780 Bochum

ACKNOWLEDGMENT

This work was supported by the DFG, in particular by the projects KI-768/4-1 and KI-768/4-2 from the ERA-chemistry, KI-768/5-1, KI-768/7-1, and KI-768/5-2 SPP-IL program. Computer time from the RZ Leipzig and NIC Jülich are gratefully acknowledged. J.F. acknowledges Prof. M. Dolg for the support in the development of the incremental scheme and for computer time which made the CC calculations in this work possible.

REFERENCES

- (1) Ludwig, R. *Angew. Chem., Int. Ed.* **2001**, *40*, 1808–1827.
- (2) Kirchner, B.; Reiher, M. *J. Am. Chem. Soc.* **2002**, *124*, 6206–6215.
- (3) Maerker, C.; v. R. Schleyer, P.; Liedl, K. R.; Ha, T. K.; Quack, M.; Suhm, M. A. *J. Comput. Chem.* **1997**, *18*, 1695–1719.
- (4) Klopper, W.; Quack, M.; Suhm, M. A. *J. Chem. Phys.* **1998**, *108*, 10096–10115.
- (5) Rincón, L.; Almeida, R.; García-Aldea, D.; Diez y Riega, H. *J. Chem. Phys.* **2001**, *114*, 5552–5561.
- (6) Salvador, P.; Szczyński, M. M. *J. Chem. Phys.* **2003**, *118*, 537–549.
- (7) Guedes, R. C.; do Couto, P. C.; Cabral, B. J. C. *J. Chem. Phys.* **2003**, *118*, 1272–1281.
- (8) Klopper, W.; Lüthi, H. P. *Mol. Phys.* **1999**, *96*, 559–570.
- (9) Vaval, N.; Kumar, A. B.; Pal, S. *Int. J. Mol. Sci.* **2001**, *2*, 89–102.
- (10) Buth, C.; Paulus, B. *Chem. Phys. Lett.* **2004**, *398*, 44–49.
- (11) Vaval, N.; Pal, S. *Chem. Phys. Lett.* **2004**, *398*, 194–200.
- (12) Buth, C.; Paulus, B. *Phys. Rev. B* **2006**, *74*, 045122.
- (13) Hirata, S.; Podeszwa, R.; Tobita, M.; Bartlett, R. J. *J. Chem. Phys.* **2004**, *120*, 2581–2592.
- (14) Sode, O.; Keçeli, M.; Hirata, S.; Yagi, K. *Int. J. Quantum Chem.* **2009**, *109*, 1928–1939.
- (15) Shiozaki, T.; Valeev, E. F.; Hirata, S. *J. Chem. Phys.* **2009**, *131*, 044118.
- (16) Izvekov, S.; Voth, G. A. *J. Phys. Chem. B* **2005**, *109*, 6553–6586.
- (17) Cournoyer, M. E.; Jorgensen, W. L. *Mol. Phys.* **1984**, *51*, 119–132.
- (18) Honda, K.; Kitaura, K.; Nishimoto, K. *Bull. Chem. Soc. Jpn.* **1992**, *65*, 3122–3134.
- (19) Jedlovsky, P.; Vallauri, R. *J. Chem. Phys.* **1997**, *107*, 10166–10176.
- (20) Valle, R. G. D.; Gazzillo, D. *Phys. Rev. B* **1999**, *59*, 13699–13706.
- (21) Quack, M.; Stohner, J.; Suhm, M. A. *J. Mol. Struct.* **2001**, *599*, 381–425.
- (22) Liem, S. Y.; Popelier, P. L. A. *J. Chem. Phys.* **2003**, *119*, 4560–4566.
- (23) Wierchowski, S. J.; Kofke, D. A. *J. Chem. Phys.* **2003**, *119*, 6092–6099.
- (24) Wierchowski, S. J.; Fang, Z. H.; Kofke, D. A.; Tilson, J. L. *Mol. Phys.* **2006**, *104*, 503–513.
- (25) Huber, H.; Dyson, A. J.; Kirchner, B. *Chem. Soc. Rev.* **1999**, *28*, 121–133.
- (26) Chen, B.; Siepmann, J. *Ilja J. Phys. Chem. B* **2000**, *104*, 8725–8734.
- (27) Deraman, M.; Dore, J.; Powles, J.; Holloway, J. H.; Chieux, P. *Mol. Phys.* **1985**, *55*, 1351–1367.
- (28) McLain, S. E.; Benmore, C. J.; Siewenie, J. E.; Urquidi, J.; Turner, J. F. C. *Angew. Chem., Int. Ed.* **2004**, *43*, 1952–1955.
- (29) Howard, B. J.; Dyke, T. R.; Klemperer, W. *J. Chem. Phys.* **1984**, *81*, 5417–5425.
- (30) Weinhold, F. *J. Chem. Phys.* **1998**, *109*, 367–372.
- (31) Weinhold, F. *J. Chem. Phys.* **1998**, *109*, 373–384.
- (32) Kirchner, B. *Phys. Rep.* **2007**, *440*, 1–111.
- (33) Pfeleiderer, T.; Waldner, I.; Bertagnolli, H.; Tölheide, K.; Fischer, H. E. *J. Chem. Phys.* **2000**, *113*, 3690–3696.
- (34) Klein, M. L.; McDonald, I. R. *J. Chem. Phys.* **1979**, *71*, 298–308.
- (35) Röthlisberger, U.; Parrinello, M. *J. Chem. Phys.* **1997**, *106*, 4658–4664.
- (36) Kreitmair, M.; Bertagnolli, H.; Mortensen, J. J.; Parrinello, M. *J. Chem. Phys.* **2003**, *118*, 3639–3645.
- (37) McGrath, M. J.; Ghogomu, J. N.; Mundy, C. J.; Kuo, I-F. W.; Siepmann, J. *Ilja Phys. Chem. Chem. Phys.* **2010**, *12*, 7678–7687.
- (38) Suhm, M. A.; Farrell, J. T.; Ashworth, S. H.; Nesbitt, D. J. *J. Chem. Phys.* **1993**, *98*, 5985–5989.
- (39) Quack, M.; Schmitt, U.; Suhm, M. A. *Chem. Phys. Lett.* **1997**, *269*, 29–38.
- (40) Haase, F.; Ahlrichs, R. *J. Comput. Chem.* **1993**, *14*, 907–912.
- (41) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165.
- (42) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.
- (43) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
- (44) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- (45) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1997**, *78*, 1396–1396.
- (46) Eichkorn, K.; Treutler, O.; Öhm, H.; Häser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *240*, 283.
- (47) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200–1211.
- (48) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822.
- (49) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553–566.
- (50) Wells, B. H.; Wilson, S. *Chem. Phys. Lett.* **1983**, *101*, 429–434.
- (51) Neugebauer, J.; Reiher, M.; Kind, C.; Hess, B. A. *J. Comput. Chem.* **2002**, *23*, 895–910.
- (52) Reyes, A.; Pak, M. V.; Hammes-Schiffer, S. *J. Chem. Phys.* **2005**, *123*, 064104.

- (53) Hurley, M.; Hammes-Schiffer, S. *J. Phys. Chem. A* **1997**, *101*, 3977–3989.
- (54) Carra, C.; Irdanova, N.; Hammes-Schiffer, S. *J. Phys. Chem. B* **2002**, *106*, 8415–8421.
- (55) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Schütz, M. et al. *MOLPRO*, version 2006; University College Cardiff Consultants Limited: Wales, U.K., 2006; <http://www.molpro.net>.
- (56) Hampel, C.; Peterson, K.; Werner, H.-J. *Chem. Phys. Lett.* **1992**, *190*, 1.
- (57) Deegan, M. J. O.; Knowles, P. J. *Chem. Phys. Lett.* **1994**, *227*, 321.
- (58) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007.
- (59) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796.
- (60) Stoll, H. *Chem. Phys. Lett.* **1992**, *191*, 548.
- (61) Friedrich, J.; Hanrath, M.; Dolg, M. *J. Chem. Phys.* **2007**, *126*, 154110.
- (62) Friedrich, J.; Hanrath, M.; Dolg, M. *J. Phys. Chem. A* **2007**, *111*, 9830.
- (63) Friedrich, J.; Dolg, M. *J. Chem. Phys.* **2008**, *129*, 244105.
- (64) Friedrich, J.; Dolg, M. *J. Chem. Theor. Comp.* **2009**, *5*, 287.
- (65) Friedrich, J.; Walczak, K.; Dolg, M. *Chem. Phys.* **2009**, *356*, 47.
- (66) Hehre, W. J.; Ditchfield, R.; Pople, J. A. *J. Chem. Phys.* **1972**, *56*, 2257.
- (67) Halkier, A.; Helgaker, T.; Jørgensen, P.; Koch, H.; Olsen, J.; Wilson, A. K. *Chem. Phys. Lett.* **1998**, *286*, 243.
- (68) Tew, D. P.; Klopper, W.; Neiss, C.; Hättig, C. *Phys. Chem. Chem. Phys.* **2007**, *9*, 1921.
- (69) Klopper, W.; Samson, C. C. M. *J. Chem. Phys.* **2002**, *116*, 6397–6410.
- (70) Peterson, K.; Adler, T.; Werner, H.-J. *J. Chem. Phys.* **2008**, *128*, 084102.
- (71) Yousaf, K. E.; Peterson, K. *J. Chem. Phys.* **2008**, *129*, 184108.
- (72) Hättig, C. *Phys. Chem. Chem. Phys.* **2005**, *7*, 59.
- (73) Ten-no, S. *J. Chem. Phys.* **2004**, *121*, 117–129.
- (74) Tew, D. P.; Klopper, W. *J. Chem. Phys.* **2005**, *123*, 074101.
- (75) Adler, T. B.; Knizia, G.; Werner, H.-J. *J. Chem. Phys.* **2007**, *127*, 221106.
- (76) Jurecka, P.; Sponer, J.; Cerny, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985.
- (77) Helgaker, T.; Jørgensen, P.; Olsen, J. *Molecular Electronic Structure Theory*; Wiley-VCH: Chichester, U.K., 2004.
- (78) Halkier, A.; Klopper, W.; Helgaker, T.; Jørgensen, P.; Taylor, P. R. *J. Chem. Phys.* **1999**, *111*, 9157.
- (79) Knizia, G.; Adler, T. B.; Werner, H.-J. *J. Chem. Phys.* **2009**, *130*, 054104.
- (80) Klopper, W.; Quack, M.; Suhm, M. S. *Chem. Phys. Lett.* **1996**, *261*, 35.
- (81) Tschumper, G. S.; Yamaguchi, Y.; Schaefer, H. F., III. *J. Chem. Phys.* **1997**, *106*, 9627.
- (82) Klopper, W.; Quack, M.; Suhm, M. S. *Mol. Phys.* **1998**, *94*, 105.
- (83) Marchetti, O.; Werner, H.-J. *J. Phys. Chem. A* **2009**, *113*, 11580.
- (84) Miller, R. E. *Acc. Chem. Res.* **1990**, *23*, 10.
- (85) Pine, A. S.; Howard, B. J. *J. Chem. Phys.* **1986**, *84*, 590–596.
- (86) Boese, A. D.; Martin, J. M. L.; Klopper, W. *J. Phys. Chem. A* **2007**, *111*, 11122.
- (87) Li, J. *J. Theor. Comp. Chem.* **2006**, *5*, 187–196.
- (88) Ruden, T. A.; Helgaker, T.; Jørgensen, P.; Olsen, J. *J. Chem. Phys.* **2004**, *121*, 5874.
- (89) Müller, H.; Franke, R.; Vogtner, S.; Jaquet, R.; Kutzelnigg, W. *Theor. Chem. Acc.* **1998**, *100*, 85–102.
- (90) Heckert, M.; Kallay, M.; Tew, D. P.; Klopper, W.; Gauss, J. *J. Chem. Phys.* **2006**, *125*, 044108.
- (91) Chaban, G. M.; Gerber, R. B. *Spectrochim. Acta, Part A* **2002**, *58*, 887–898.
- (92) Karpfen, A.; Yanovitskii, O. *J. Mol. Struct. (THEOCHEM)* **1994**, *314*, 211–227.
- (93) Karpfen, A. *Chem. Phys.* **1980**, *47*, 401–406.
- (94) Karpfen, A.; Schuster, P. *Chem. Phys. Lett.* **1976**, *44*, 459–464.
- (95) Huber, K. P.; Herzberg, G. *Molecular Spectra and Molecular Structure. IV Constants of Diatomic Molecules*; Van Nostrand: New York, 1979.
- (96) Tew, D. P.; Klopper, W.; Heckert, M.; Gauss, J. *J. Phys. Chem. A* **2007**, *111*, 11242.
- (97) Swalina, C.; Wang, Q.; Chakraborty, A.; Hammes-Schiffer, S. *J. Phys. Chem. A* **2007**, *111*, 2206–2212.
- (98) Martin, J. M. L.; Taylor, P. R. *Chem. Phys. Lett.* **1994**, *225*, 473.
- (99) Feller, D.; Peterson, K. A. *J. Mol. Struct. (THEOCHEM)* **1997**, *400*, 69.

Polarized Molecular Orbital Model Chemistry. 1. Ab Initio Foundations

Luke Fiedler, Jiali Gao,* and Donald G. Truhlar*

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

 Supporting Information

ABSTRACT: The objective of this paper is to examine the minimal requirements for obtaining semiquantitative polarizabilities of molecules, in order to provide a well-founded starting point for a new semiempirical molecular orbital formulation that is more suitable than presently available methods for simulating electronic polarization effects. For this purpose, we present polarizability calculations for 38 molecules with 36 basis sets, including many unconventional ones, and five semiempirical molecular orbital theories based on neglect of diatomic differential overlap. We conclude that two basis sets are particularly promising to serve as bases for semiempirical improvement, namely, STO-3G(P), in which diffuse p functions are added to all hydrogens, and 3-(21,3,21)G, in which a minimal basis set is augmented with one extra s function on every atom. We especially recommend the former because all intra-atomic overlap integrals are zero by symmetry, which makes it a better candidate for neglect-of-differential-overlap treatments.

1. INTRODUCTION

Semiempirical molecular orbital theory with the neglect of diatomic differential overlap¹ (NDDO, e.g., MNDO,² AM1,³ PM3,⁴ and RM1⁵) has been enormously useful. It continues to be widely used in applications to macromolecular systems, although it is not without known deficiencies. One of the chief deficiencies is that it does not describe intermolecular interactions very accurately, partially due to underestimation of molecular polarizabilities.^{6,7} This deficiency is a major limiting factor for quantitative simulation and modeling of systems of biological and materials interest. As an example of trying to remedy the underpolarization of NDDO methods, Schürer et al.⁸ reported a parametrized variational method for calculating molecular electronic polarizabilities from an NDDO wave function. That method, however, does not concern the polarization of the electronic wave function itself, and hence it is unsuitable for modeling polarization in direct dynamics calculations. Our goal is to include polarization in the model wave function itself to provide a framework for developing a new generation of force fields for large molecules in condensed phases, where the new force fields go beyond fixed-charge molecular mechanics by allowing for self-consistent polarization and charge transfer.⁹

One can make up for “small” qualitative deficiencies in molecular model chemistries by parametrization if one uses an appropriate theoretical framework, but it is hard and dangerous to use parametrization to make up for large quantitative deficiencies or for a framework that does not contain the dominant physical factors. In the latter case, it is desirable to improve the form of the model so the deficiencies that must be overcome are smaller and the model to be parametrized contains the physical features that correctly account for the phenomenon to be modeled.

Current NDDO molecular orbital theory for elements in the 1s, 2s, and 2p blocks of the periodic table usually has the form of Hartree–Fock (HF) theory employing a minimum basis set

(MBS) with many integrals systematically neglected and others parametrized rather than evaluated as in *ab initio* calculations.¹ Since the parametrization is carried out against experimental data rather than against a correct HF/MBS calculation, not only does it make up for the neglected integrals but it also empirically introduces higher-order effects, in particular, some of the effects of using a more complete (more polarizable) basis set and some of the effects of electron correlation. Using the STO-3G¹⁰ minimum basis set, *ab initio* Hartree–Fock theory underestimates the mean polarizability of a water molecule by a factor of 3.6, which may be compared to a factor of 3.4 for MNDO and a factor of 2.9 for AM1, PM3, and RM1. Thus, these semiempirical parametrizations only improved the polarizability marginally. When we tried to reparameterize these models to have greater polarizability, many of the other properties became significantly worse, which indicates a fundamental limitation in building the model on the HF/MBS framework. An interesting question is whether there is an *ab initio* formalism, not much more complicated than HF/MBS, that has a significantly reduced error in computed molecular polarizability, e.g., less than a factor of 2 (as compared to a factor of 3.4). An NDDO version of such a theory might prove easier to parametrize satisfactorily than the NDDO version of HF/MBS theory. For example, should we explicitly include electron correlation? We already know from accumulated experience with *ab initio* calculations that this is not necessary; i.e., large-basis-set HF theory gives realistic (although of course not quantitatively accurate) polarizabilities. But how much larger a basis set is called for? What is the smallest increase in basis set that yields qualitatively correct polarizabilities? The present article is devoted to this question. Part II¹¹ builds on the findings of this paper to develop a new NDDO framework, which we call the polarized molecular orbital (PMO) model, that yields reasonable results for polarizabilities.

Received: November 7, 2010

Published: March 03, 2011

Section II of the present paper presents a database that will be used to test basis sets for polarizabilities. Section III then presents Hartree–Fock polarizability calculations with 36 different basis sets.

II. DATABASE

To understand the basis set requirements for predicting realistic polarizabilities, we consider only mean dipole polarizabilities, α , defined as

$$\alpha = \frac{\alpha_{xx} + \alpha_{yy} + \alpha_{zz}}{3} \quad (1)$$

where α_{ij} is an element of the electric dipole polarizability tensor. The mean polarizability is one-third of the trace of the polarizability tensor and is independent of the orientation of the axes. In the rest of this article and the next one, we will simply call α the polarizability.

Table 1 gives the 38 polarizabilities that we will use to test basis sets. In most cases, these are taken from experimental results;^{12–21} in a few cases, where we did not have experimental values, we calculated the polarizability using MP2/aug-cc-pVTZ//MP2/aug-cc-pVDZ, MP2/aug-cc-pVDZ, or HF/aug-cc-pVDZ, where HF denotes Hartree–Fock, MP2 denotes Møller–Plesset second-order perturbation theory,²² and we use the usual notation²³ for correlation-consistent basis sets.

III. BASIS SETS

The basis sets considered are in Table 2. For most of the basis sets, the notation is standard. Nonstandard basis sets are explained in the table; in these explanations, exponential parameters are called exponents and are given in a_0^{-2} . All basis sets are single- ζ for the core orbitals; for valence electrons, we consider both regular valence basis functions (indicated by lower case s, p, or d) and diffuse basis functions (indicated by capital S, P, or D). The table shows how many basis functions of each type (excluding core basis functions) are present in a given basis: N_1 is the number of basis functions on each hydrogenic atom, and N_2 is the number of basis functions (excluding the core) on each nonhydrogenic atom. (To obtain the total number of basis functions, add one core basis function for C, N, and O and five core basis functions for S.)

IV. RESULTS

Polarizabilities for three typical molecules are given in Table 3. We give results for all 36 basis sets of Table 2 plus five NDDO methods: MNDO,² AM1,³ PM3,⁴ RM1,⁵ and PM6.²⁴ A complete table of polarizability calculations for all 38 molecules is presented in the Supporting Information. Table 3 also gives the mean unsigned percentage error (MUPE, that is, the mean of the absolute values of all 38 percentage deviations of the calculated polarizabilities from the reference values of Table 1). The polarizability tensor components of two molecules, water and acetamide, are given in Table 4.

V. DISCUSSION

The final five rows of Table 3 confirm that presently available semiempirical methods underestimate molecular polarizabilities, but comparison to the results using STO-3G shows that they have smaller errors than *ab initio* Hartree–Fock calculations with the same size basis set. The other results in the table explore the effect of expanding the basis set.

Table 1. Polarizabilities (\AA^3)

molecule	α (\AA^3)	ref
H ₂ O	1.45	12b
CH ₄	2.59	19
HCN	2.59	13
HCl	2.63	13
(H ₂ O) ₂	2.88	a
CO ₂	2.91	16
CH ₃ OH	3.23	12a
CHCH	3.33	13
CH ₃ NH ₂	4.01	14
CH ₂ CH ₂	4.25	17
CH ₃ CH ₃	4.48	13
CH ₃ CN	4.48	18
CH ₃ CHO	4.59	18
CH ₃ Cl	4.72	15
CH ₂ CHOH	4.85	a
CH ₃ CH ₂ OH	5.11	18
CH ₃ OCH ₃	5.16	12a
CH ₃ SH	5.49	a
CH ₃ C(O)NH ₂	5.67	18
CHCCH ₃	6.18	12b
CH ₂ CHCH ₃	6.26	12b
CH ₃ C(O)CH ₃	6.33	12a
CH ₃ CH ₂ CH ₃	6.38	12b
H ₂ NCH ₂ COOH	6.52	a
CH ₃ C(O)Cl	6.62	12b
CH ₃ CHOHCH ₃	7.61	12b
CH ₃ SCH ₃	7.39	a
pyrimidine	8.53	20
<i>s-trans</i> -butadiene	8.64	12b
pyridine	9.18	21
diethyl amine	9.61	21
benzene	10.32	13
phenol	11.10	12b
toluene	12.26	12a
nicotinamide	12.19	b
purine	12.78	c
benzaldehyde	12.80	c
benzyl alcohol	13.15	c

^a Present work, calculated by MP2/aug-cc-pVTZ//MP2/aug-cc-pVDZ.

^b Present work, calculated by HF/aug-cc-pVDZ. ^c Present work, calculated by MP2/aug-cc-pVDZ.

The table shows that adding diffuse P functions on hydrogen atoms is particularly effective in increasing the polarizabilities. For example, the STO-3G(P) basis set has a smaller MUPE than the much larger STO-3G(d), 3-21G, 3-21G(d), or 3-21G(p) basis sets, and the error is almost as small as for the well-polarized cc-pVDZ basis set or the 3-21+G(d) basis set, which contains more contracted functions than the cc-pVDZ basis set. (We note that extending NDDO methodology to the 4-31G basis set, which is similar to the 3-21G basis set, has already been considered by Thiel;²⁵ however, results for polarizabilities were not presented.)

Deleting the diffuse P functions on any hydrogens, even deleting them only on hydrogens attached to sp³ carbon atoms, raises the MUPE appreciably, as shown by the results for STO-3G(P*).

Table 2. Basis Sets and Numbers of Regular Valence Basis Functions and Diffuse Basis Functions

basis	note	non-hydrogenic					hydrogenic							
		s	p	d	S	P	D	N ₁	s	p	S	P	N ₂	N ₁ +N ₂
STO-3G		1	3					4	1				1	5
STO-3G+		1	3		1	3		8	1				1	9
STO-3G++		1	3		1	3		8	1	1			2	10
STO-3G(P)	<i>a</i>	1	3					4	1		3	4		8
STO-3G+(P)	<i>a</i>	1	3		1	3		8	1		3	4		12
STO-3G++(P)	<i>a</i>	1	3		1	3		8	1	1	3	5		13
STO-3G(p)	<i>b</i>	1	3					4	1	3			4	8
STO-3G(pP)	<i>a, b</i>	1	3					4	1	3	3	7		11
STO-3G(D,P)	<i>a, c</i>	1	3				5	9	1		3	4		13
STO-3G(P,P)	<i>a, c</i>	1	3			3		7	1		3	4		11
STO-3G(S,P)	<i>a, c</i>	1	3		1			5	1		3	4		9
3-21G		2	6					8	2				2	10
3-21+G		2	6		1	3		12	2				2	14
3-21++G		2	6		1	3		12	2	1			3	15
3-21G(P)	<i>a</i>	2	6					8	2		3	5		13
3-21+G(P)	<i>a, c</i>	2	6		1	3		12	2		3	5		17
3-21++G(P)	<i>a, c</i>	2	6		1	3		12	2	1	3	6		18
3-21G(p)	<i>b</i>	2	6					8	2	3			5	13
3-21G(pP)	<i>a, b</i>	2	6					8	2	3	3	8		16
3-21G(D,P)	<i>a, c</i>	2	6				5	13	2		3	5		18
STO-3G(d)	<i>d</i>	1	3	5				9	1				1	10
3-21G(d)	<i>d</i>	2	6	5				13	2				2	15
3-21+G(d)	<i>d</i>	2	6	5	1	3		17	2				2	19
3-(21,3,3)G	<i>e</i>	2	3					5	1				1	6
cc-pVDZ		2	6	5				13	2	3			5	18
aug-cc-pVDZ		2	6	5	1	3	5	22	2	3	1	3	9	31
3-(21,21,3)G	<i>e</i>	2	6					8	1				1	9
3-(21,3,21)G	<i>e</i>	2	3					5	2				2	7
3-(21,3,3)G(P)	<i>a, e</i>	2	3					5	1		3	4		9
3-(3,21,21)G	<i>e</i>	1	6					7	2				2	9
aug'-cc-pVDZ		2	6	5	1	3	5	22	2	3	1		6	28
aug''-cc-pVDZ		2	6		1	3		12	2		1		3	15
STO-3G(P!)	<i>f</i>	1	3					4	1		(3)	1, 4		5, 8
STO-3G(P*)	<i>g</i>	1	3					4	1		(3)	1. 4		5, 8
STO-3G(P')	<i>h</i>	1	3					4	1			3	1	8
STO-3G(P'')	<i>i</i>	1	3					4	1			3	1	8

^a Exponent on hydrogen P orbital = 0.141, taken from aug-cc-pVDZ. Note that we use capital S, P, or D to denote diffuse basis functions, thereby distinguishing them from regular valence basis functions (indicated by lower case s, p, or d). ^b Exponent on hydrogen p orbital = 0.727, taken from cc-pVDZ. ^c Exponents for diffuse functions on non-hydrogenic atoms taken from aug-cc-pVDZ. ^d Exponent for non-hydrogenic d functions = 0.8. ^e Combination of STO-3G and 3-21G. In 3-(x,y,z)G, x denotes valence basis for nonhydrogenic s orbitals, y denotes basis for nonhydrogenic p orbitals, and z denotes basis for hydrogen. ^f P functions only on polar hydrogens, i.e., hydrogens not bonded to carbon. ^g P functions only on polar hydrogens and hydrogens bonded to non sp³ carbons. ^h Exponent on hydrogen P orbital = 0.123, which was obtained by optimizing it with respect to the polarizabilities of the 38-molecule test set. ⁱ Exponent on hydrogen P orbital = 0.082.

The introduction of p orbitals on hydrogen atoms was previously considered in semiempirical molecular orbital theory most thoroughly by Jug and Geudtner,²⁶ who added p orbitals to hydrogen atoms in the SINDO1 approximation; their goal was to

Table 3. Polarizabilities and Mean Unsigned Percentage Errors in Polarizabilities (Å³)

basis	N ₁ + N ₂	acetaldehyde	dimethyl sulfide	vinyl alcohol	MUPE
reference ^a		4.59	7.39	4.85	0 ^b
STO-3G	5	1.76	2.63	1.96	62
STO-3G+	9	3.06	5.29	3.68	32
STO-3G++	10	3.08	5.61	3.78	29
STO-3G(P)	8	3.15	5.17	3.46	32
STO-3G+(P)	12	3.93	6.83	4.47	14
STO-3G++(P)	13	3.92	6.95	4.49	14
STO-3G(p)	8	2.03	3.06	2.22	56
STO-3G(pP)	11	3.23	5.31	3.55	31
STO-3G(D,P)	13	3.59	5.98	3.97	21
STO-3G(P,P)	11	3.73	6.26	4.09	21
STO-3G(S,P)	9	3.55	5.69	3.96	24
3-21G	10	3.11	5.10	3.07	36
3-21+G	14	3.64	5.78	3.88	23
3-21++G	15	3.69	5.92	3.93	22
3-21G(P)	13	3.50	5.88	3.67	25
3-21+G(P)	17	3.94	6.46	4.32	15
3-21++G(P)	18	3.96	6.54	4.33	14
3-21G(p)	13	3.23	5.32	3.21	33
3-21G(pP)	16	3.55	5.95	3.73	24
3-21G(D,P)	18	3.86	6.40	4.21	15
STO-3G(d)	10	1.95	2.87	2.16	57
3-21G(d)	15	N/A	N/A	3.18	33
3-21+G(d)	19	3.70	N/A	4.00	20
3-(21,3,3)G	6	1.94	5.10	2.16	54
cc-pVDZ	18	3.49	5.58	3.57	26
aug-cc-pVDZ	31	4.18	6.86	4.59	8
3-(21,21,3)G	9	2.43	4.05	2.59	47
3-(21,3,21)G	7	3.51	5.06	3.35	29
3-(21,3,3)G(P)	9	4.34	6.47	4.45	11
3-(3,21,21)G	9	4.22	5.95	3.92	21
aug'-cc-pVDZ	28	4.13	6.77	4.53	9
aug''-cc-pVDZ	15	3.71	5.91	3.97	21
STO-3G(P!)	5, 8	1.76	2.63	2.18	58
STO-3G(P*)	5, 8	2.20	2.63	3.46	48
STO-3G(P')	8	2.99	5.23	3.59	32
STO-3G(P'')	8	3.13	5.22	3.50	32
MNDO	5	2.72	3.94	3.04	42
AM1	5	2.83	4.31	3.09	41
PM3	5	2.60	3.96	2.87	44
RM1	5	2.70	4.36	3.01	42
PM6	5	2.03	3.24	2.35	56

^a From Table 1. ^b By definition.

improve the treatment of hydrogen bonding. In the present work, it was found, through systematic investigation of basis set dependence, that adding P functions (that is, in the notation established in section III, diffuse p functions) on hydrogen to a minimum basis set in *ab initio* Hartree–Fock calculations provides a powerful strategy to calculate more accurate polarizabilities. The mean unsigned percentage errors is 32% for adding a P subshell but 56% for adding a p subshell. Adding both further reduces the error, but only to 31%, showing that the diffuse P

Table 4. Polarizability Tensor Components of Water and Acetamide (\AA^3)

method/basis	α_{xx}	α_{xy}	α_{yy}	α_{xz}	α_{yz}	α_{zz}
water						
HF/STO-3G	0.01	0.00	0.82	0.00	0.00	0.38
HF/STO-3G(P)	0.34	0.00	1.40	0.00	0.00	0.70
HF/3-(21,3,21)G	0.25	0.00	1.28	0.00	0.00	0.86
M06-2X/6-31+G(d,p)	0.91	0.00	1.18	0.00	0.00	1.01
HF/aug-cc-pVDZ	1.11	0.00	1.42	0.00	0.00	1.27
HF/aug-cc-pVTZ	1.16	0.00	1.44	0.00	0.00	1.32
MP2/aug-cc-pVDZ	1.33	0.00	1.53	0.00	0.00	1.42
MP2/aug-cc-pVTZ	1.40	0.00	1.55	0.00	0.00	1.48
acetamide						
HF/STO-3G	2.62	0.09	3.19	0.05	-0.14	1.29
HF/STO-3G(P)	4.65	0.14	4.64	0.14	-0.26	2.63
HF/3-(21,3,21)G	6.19	-0.25	5.46	0.13	-0.26	2.68
M06-2X/6-31+G(d,p)	5.69	-0.05	6.26	-0.02	0.03	4.05
HF/aug-cc-pVDZ	5.81	0.00	6.09	0.01	-0.01	4.27
HF/aug-cc-pVTZ	5.85	-0.01	6.10	0.01	-0.01	4.30
MP2/aug-cc-pVDZ	6.43	-0.11	6.69	-0.01	0.04	4.58
MP2/aug-cc-pVTZ	6.47	-0.12	6.70	-0.01	0.03	4.60

function is the key to the success of this strategy. One obtains a similar mean error, in particular 33%, with either the 3-21G(p) or 3-21G(d) basis set. Taking acetylene as an example, Table 2 shows that the number of valence basis functions in the STO-3G(P) basis set is 16, whereas the number in the 3-21G(p) or 3-21G(d) is 26–30. For a water molecule, these numbers are reduced to 12 for STO-3G(P) and to 17–18 for 3-21G(p) or 3-21G(d).

Another strategy that may be compared to the successful STO-3G(P) strategy is to add a d subshell to the STO-3G basis to every nonhydrogenic atom, yielding the STO-3G(d) basis set. This might have been anticipated to be a powerful strategy (the natural choice) because most of the electrons in the molecules considered here are in valence p orbitals, and these are strongly coupled to d functions by the electric dipole operator. Using this strategy raises the number of valence basis functions to 20 for acetylene and lowers it to 11 for water, but it yields a mean error of 57%. Thus, this strategy is less successful.

A second successful strategy revealed by Tables 2 and 3 is to split the valence s subshell of STO-3G on all atoms. If one were to split all of the valence subshells, that would give 3-21G, but splitting only s subshells gives 3-(21,3,21)G. The mean error is 29%, and the number of valence basis functions for the example case of acetylene is only 14. For a water molecule, this number is reduced to 9. Of the two successful strategies, i.e., (a) splitting the valence s subshell on all atoms and (b) adding a set of diffuse P functions to hydrogen atoms, the latter is preferred for two reasons. First, the inclusion of a set of P orbitals allows the out-of-plane polarizability as well as the in-plane polarizability to be better represented, which is particularly important for key compounds such as water and benzene or any other planar molecule. Second, when the strategy is applied in the context of the NDDO approximation, choice b avoids the treatment of nonorthogonal basis functions on the same center, which is probably very important since neglect of one-center differential overlap has been singled out as a significant shortcoming of the current NDDO formalisms, even with minimal basis sets.^{27,28} We examine the

use of the STO-3G(P) basis as a starting point for NDDO parametrization in the following article.

The polarizability tensor components for water and acetamide are provided in Table 4, with axes aligned along the principal axes. For comparison, we also give results calculated with the same axis choices with the M06-2X density functional²⁹ and ab initio MP2 wave function theory²² (with standard basis sets^{23,30}). Water is in the yz plane, and acetamide has the yz plane as a plane of C_s symmetry. By comparing the results with the STO-3G basis set to those with the STO-3G(P) and 3-(21,3,21)G basis sets, the table shows that both new basis sets allow out-of-plane polarization, and for the larger molecule, the error in out-of-plane polarizability is less than 30%.

VI. CONCLUSIONS

We found that adding only a few well chosen extra basis functions can considerably reduce the errors in minimum-basis-set calculations of electric dipole polarizabilities. Two particularly efficient strategies are (i) to add a diffuse P subshell to every hydrogen atom and (ii) to split the valence s shell on each atom into a double- ζ pair of s functions. It is particularly encouraging that these strategies work better than the seemingly more obvious choice of adding d functions to nonhydrogenic atoms.

■ ASSOCIATED CONTENT

S Supporting Information. A complete set of tables of polarizability calculations and geometries optimized at the MP2 level. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: gao@jialigao.org (J.G.) and truhlar@umn.edu (D.G.T.).

■ ACKNOWLEDGMENT

This work was supported in part by the National Institutes of Health (grant no. RC1-GM091445) and the National Science Foundation (grant no. CHE09-56776).

■ REFERENCES

- (1) Pople, J. A.; Santry, D. P.; Segal, G. A. *J. Chem. Phys.* **1965**, *43*, S129.
- (2) Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4899.
- (3) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
- (4) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 209.
- (5) Rocha, G. B.; Freire, R. O.; Simos, A. M.; Stewart, J. J. P. *J. Comput. Chem.* **2006**, *27*, 1101.
- (6) Dewar, M. J. S.; Yamaguchi, Y.; Suck, S. *Chem. Phys. Lett.* **1974**, *59*, 541.
- (7) Parkinson, W. A.; Zerner, M. C. *J. Chem. Phys.* **1991**, *94*, 478.
- (8) Schürer, G.; Geddeck, P.; Gottschalk, M.; Clark, T. *Int. J. Quantum Chem.* **1999**, *75*, 17.
- (9) (a) Gao, J. *J. Phys. Chem. B* **1997**, *101*, 657. (b) Gao, J. *J. Chem. Phys.* **1998**, *109*, 2346. (c) Xie, W.; Gao, J. *J. Chem. Theory Comput.* **2007**, *3*, 1890. (d) Xie, W.; Song, L.; Truhlar, D. G.; Gao, J. *J. Chem. Phys.* **2008**, *128*, 234108. (e) Xie, W.; Song, L.; Truhlar, D. G.; Gao, J. *J. Phys. Chem. B* **2008**, *112*, 14124. (f) Song, L.; Han, J.; Lin, Y. L.; Xie, W.; Gao, J. *J. Phys. Chem. A* **2009**, *113*, 11656. (g) Xie, W.; Orozco, M.; Truhlar, D. G.; Gao, J. *J. Chem. Theory Comput.* **2009**, *5*, 459. (h) Cembran, A.; Bao, P.; Wang, Y.; Song, L.; Truhlar, D. G.; Gao, J. Work in progress.

- (10) Hehre, W. J.; Stewart, R. F.; Pople, J. A. *J. Chem. Phys.* **1969**, *51*, 2657.
- (11) Zheng, P.; Fiedler, L.; Leverentz, H.; Truhlar, D. G.; Gao, J. *J. Chem. Theory Comput.* **2011**; DOI: 10.1021/ct100638g.
- (12) (a) Stuart, H. A. In *Landolt-Börnstein "Zahlenwerte und Funktionen"*; Eucken, A., Hellwege, K. H., Eds.; Springer-Verlag: Berlin, 1951; p 511. (b) Maryott, A. A.; Buckley, F. *U. S. National Bureau of Standards Circular No. 537*; U. S. National Bureau of Standards: Washington, DC, 1953.
- (13) Hirschfelder, J. O.; Curtiss, C. F.; Bird, R. B. *Molecular Theory of Gases and Liquids*; Wiley: New York, 1954; p 950.
- (14) Bridge, N. J.; Buckingham, A. D. *Proc. R. Soc. (London)* **1966**, *295*, 334.
- (15) Sutter, H.; Cole, R. H. *J. Chem. Phys.* **1970**, *52*, 132.
- (16) Bose, T. K.; Cole, R. H. *J. Chem. Phys.* **1970**, *52*, 140.
- (17) Bose, T. K.; Cole, R. H. *J. Chem. Phys.* **1971**, *54*, 132.
- (18) Applequist, J.; Carl, J. R.; Fung, K.-K. *J. Am. Chem. Soc.* **1972**, *94*, 2952.
- (19) Bose, T. K.; Sochanski, J. S.; Cole, R. H. *J. Chem. Phys.* **1972**, *57*, 3592.
- (20) Marchese, F. T.; Jaffe, H. H. *Theor. Chim. Acta* **1977**, *45*, 241.
- (21) No, K. T.; Cho, K. H.; Jhon, M. S.; Scheraga, H. A. *J. Am. Chem. Soc.* **1993**, *115*, 2005.
- (22) Pople, J. A.; Binkley, J. S.; Seeger, R. *Int. J. Quantum Chem. Symp.* **1976**, *10*, 1.
- (23) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796.
- (24) Stewart, J. J. P. *J. Mol. Model* **2007**, *13*, 1173.
- (25) Thiel, W. *Theor. Chim. Acta* **1981**, *59*, 191.
- (26) Jug, K.; Geudtner, G. *J. Comput. Chem.* **1993**, *14*, 639.
- (27) Thiel, W. *Theor. Chem. Acc.* **2000**, *103*, 495.
- (28) Winget, P.; Selcuki, C.; Horn, A. H. C.; Martin, B.; Clark, T. *Theor. Chem. Acc.* **2003**, *110*, 254.
- (29) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215.
- (30) Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; Schleyer, P. v. R. *J. Comput. Chem.* **1983**, *4*, 294.

Polarized Molecular Orbital Model Chemistry. 2. The PMO Method

Peng Zhang, Luke Fiedler, Hannah R. Leverentz, Donald G. Truhlar,* and Jiali Gao*

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

S Supporting Information

ABSTRACT: We present a new semiempirical molecular orbital method based on neglect of diatomic differential overlap. This method differs from previous NDDO-based methods in that we include p orbitals on hydrogen atoms to provide a more realistic modeling of polarizability. As in AM1-D and PM3-D, we also include damped dispersion. The formalism is based on the original MNDO one, but in the process of parametrization we make some specific changes to some of the functional forms. The present article is a demonstration of the capability of the new approach, and it presents a successful parametrization for compounds composed only of hydrogen and oxygen atoms, including the important case of water clusters.

1. INTRODUCTION

Molecular modeling methods for calculating potential energy surfaces and forces span a wide spectrum of accuracy and cost, ranging from pairwise potentials in molecular mechanics on the one hand to coupled cluster theory at the complete basis set limit on the other. But there is a large gap in computational cost between using semiempirical molecular orbital methods at the high end of the low end and small-basis-set Hartree–Fock or density functional theory at the low end of the high end. The goal of this work is to develop a new method in that gap, which can be applied to the study of the internal energies of a single molecule or a group of molecules and to macromolecular systems, including solvated proteins. In particular, we seek a method that is almost as inexpensive as the popular AM1¹ and PM3² semiempirical methods (both based on neglect of diatomic differential overlap^{3,4} (NDDO)) but is more accurate in two key respects: (1) it gives more accurate noncovalent interactions, and (2) it yields more accurate molecular polarizabilities. Such a method would be well suited for use in the Explicit Polarization^{5,6} (X-Pol) force field or in any simulation where it is important to include polarization and induction effects in a large system that requires extensive sampling of fluctuations or conformational states.

Our starting point is the MNDO⁴ formalism, which has been the basis for the later, more successful AM1 and PM3 through PM6 series of model chemistries. It has the advantage as a starting point for new work in the treatment of effective nuclear core–core interactions. To this end, we make three key enhancements to MNDO:

- A set of p-type basis functions is added on the hydrogen atoms. The motivation for this is provided in the previous paper,⁷ and we also note (i) the work of Parkinson and Zerner,⁸ adding p functions on hydrogen to the INDO model chemistry to improve the calculation of hyperpolarizabilities and (ii) the SINDO1 and MSINDO methods of Jug and Bredow in which the basis can be augmented with p functions, where the motivation was a better treatment of hydrogen bonding.⁹
- A damped dispersion function is included as in the PM3-D method of Hillier and co-workers.^{10–12}

- The new approximate molecular orbital method is parametrized against molecular polarizabilities and noncovalent interactions as well as static molecular data.

Since explicit p-type polarization functions on hydrogen atoms are incorporated into our method, it is called the polarized molecular orbital (PMO) model. We envision developing a new-generation semiempirical model that can be applied equally well both to the study of the internal energies of a single molecule or cluster of molecules and to dynamical simulation and modeling of condensed-phase systems, including liquids, liquid-phase solutions, and biological macromolecular interactions. Since water is essential for applications to all biomolecular interactions and to many other chemical systems, we first focus on the development of a model for compounds composed of hydrogen and oxygen that can be used as a basic starting point to create a balanced PMO model for describing intermolecular interactions. In particular, in the present article, we present the key theoretical ingredients, parametrization strategies, and computational details of PMO by parametrizing it for molecules composed of only O and H atoms. In this process, we make some changes in the functional forms used for fitting electronic integrals in MNDO, but we consider these parametrization choices, not essential parts of the method. In principle, they could be changed in the future to develop an even more general model to encompass all functionalities of chemical and biological interest. Since such future parametrizations and extensions to other elements are anticipated, the present parametrization of PMO is labeled as version 1 or PMOv1 in our computer program. However, since it is the only version currently available, no confusion will occur if we simply call it PMO in the present article.

2. THEORY

As in all popular semiempirical models, only valence electrons are treated explicitly; the nuclei and core electrons (for atoms

Received: November 8, 2010

Published: March 03, 2011

heavier than He) are combined together as the nuclear “core.” The core–core interactions are used to describe classical Coulomb energies and to account for the electronic screening effects and the errors in electronic integrals introduced by the NDDO approximation. The core–core repulsion is modeled in an empirical way, even for the core of a hydrogen atom, which is just a nucleus. Valence electrons are treated by a minimum basis set of one *s* function and a subshell of three *p* functions on each atom; molecular orbitals are optimized by an iterative self-consistent-field¹³ (SCF) calculation.

In principle, semiempirical models are parametrized against experimental data, so that electron correlation effects, including short- and medium-range dynamical correlation effects and long-range dispersion interactions, can be implicitly incorporated into the method by using effective electronic integrals. This has been the strategy employed by Dewar¹ and others in the past; however, significant progress in the understanding of the origin of intermolecular interactions has been made in the past 25 years, especially by using accurate ab initio methods. The relative contribution of dispersion effects to intermolecular interactions can now be accurately estimated for small- to medium-sized systems. A number of recent methods for energy decomposition analysis further helped the understanding of the origin of binding interactions in molecular complexes.^{14–17} Thus, we have decided to incorporate empirical dispersion terms as a post-SCF procedure, and in our current implementation, the same functional form is adopted as that used by several groups previously,^{10–12,18,19} with the same parameters as used by Grimme¹⁹ and by Hillier and co-workers.^{10–12}

Another difference in the present treatment from previous NDDO parametrizations is that we parametrize the electronic energy, including nuclear repulsion, against corresponding benchmark data instead of parametrizing against standard-state enthalpies of formation.

In the NDDO approximation, interatomic differential overlap is neglected in two-electron integrals and in the overlap integrals, but all one-electron integrals are retained, although one-electron as well as two-electron integrals are parametrized rather than evaluated directly.³

In MNDO and all other subsequent versions of semiempirical models based on NDDO approximations considered here, two-center electron repulsion integrals are treated as multipole–multipole interactions by the method of Dewar and Thiel.²⁰ This is also used in PMO except for two-center terms involving *p* orbitals of hydrogen on both centers. In the MNDO treatment of two-center electron repulsion integrals, the density of a *p* orbital is represented as the sum of a monopole and a quadrupole charge distribution. In PMO, for integrals involving hydrogen *p* orbitals on both centers, we retain only the monopole contribution.

Parameters in NDDO methods depend on the atomic numbers of the orbitals involved in the parametrized integral. Because of the NDDO approximation, all three-center and four-center integrals are neglected. For the highest accuracy, all parameters for two-center integrals would be specific to a given pair of atomic numbers (pair parameter scheme, in the language used in ref 4), but in the present paper, as in the original MNDO,⁴ we generally use combining rules to obtain two-center parameters from atomic parameters that are independent of the other center (atomic-parameter scheme). For completeness, we summarize below the conventions used for atomic parameters in the MNDO method, after which, we highlight

three exceptions where we override the use of atomic parameters by using pair parameters.

There are 12 independent empirical parameters for each atomic number:

- U_{SS}^A and U_{PP}^A : the one-center, one-electron energies representing the kinetic energy of an *s* or *p* electron on center *A* and its attraction to its own core³
- β_s^A and β_p^A : atomic parameters responsible for the two-center, one-electron resonance integrals for *l* = 0 (*s*) and *l* = 1 (*p*). The resonance integrals are given by⁴

$$\beta_{ll'}^{AA'} = \frac{\beta_l^A + \beta_{l'}^{A'}}{2} \langle Alm_l | A'l'm'_l \rangle \quad (1)$$

where $|A lm_l\rangle$ denotes an atomic basis function on atom *A*, *l* and *m*_{*l*} are the angular momentum and magnetic quantum numbers of the electron on the other center *A'*, and $\langle Alm_l | A'l'm'_l \rangle$ is an overlap integral.⁴

- ζ_s^A and ζ_p^A : parameters for the exponent in the Slater-type orbitals (STOs) for atom *A* used to compute the overlap integrals of eq 1 as well as to determine the charge separations in the multipole expansion sites used to evaluate two-center, two-electron integrals.⁴ To make the gradients analytic, the STOs are expressed as a linear combination of six Gaussians.²¹
- α^A : electronic screening parameters for the exponent in the core–core repulsion terms.⁴
- g_{ss}^A , g_{sp}^A , g_{pp}^A , $g_{pp'}^A$, and h_{sp}^A : parameters used to establish the one-center limits of the Coulomb (*g*) and exchange (*h*) integrals as $R_{AA'} \rightarrow 0$ of the Dewar–Sabelli–Klopman approximation to the two-center, two-electron integrals, where $R_{AA'}$ is the internuclear distance between atoms *A* and *A'*.²² Another exchange integral is also needed, but in order to maintain rotational invariance, it is given by

$$h_{pp'}^A = \frac{g_{pp}^A + g_{pp'}^A}{2} \quad (2)$$

As mentioned above, in three cases, exceptions were made to using the standard MNDO functional forms on the basis of atomic parameters:

- (1) For the resonance integrals involving *p* orbitals on hydrogen atoms, we override eq 1 with

$$\beta_{pp}^{HH'} = 0 \quad (3a)$$

$$\beta_{sp}^{HH'} = 0 \quad (3b)$$

$$\beta_{lp}^{OH} = P_1 \frac{\beta_l^O + \beta_p^H}{2} \langle Olm_l | H1m'_l \rangle \quad (4)$$

where *l*' = 1 and P_1 is a parameter. Furthermore, for the H–H resonance integrals, we use a specific value ζ_l^{HH} of ζ_l^H , and for the O–O resonance integrals, we use a specific value ζ_l^{OO} of ζ_l^O .

- (2) For the two-center, one-electron attraction integrals between an electron in a distribution on atom H and the core of another hydrogen atom *B* (where *B* ≠ H but $Z_B = 1$), for which we use the ab initio notation $\langle Hlm_l | -Z_B/r_B | H1m'_l \rangle$, in which *l* is 1 and r_B is the distance of the electron on H from *B*, the standard MNDO result is

Table 1. Parameters

parameter	PMO	MNDO	AM1	PM3	PDDG/PM3	RM1	PM3-D	PM6
U_{ss}^H (eV)	-11.22813	-11.91	-11.40	-13.07	-12.89	-11.96	-13.05	-11.25
U_{pp}^H (eV)	-9.95254							
β_s^H (eV)	-6.89857	-6.99	-6.17	-5.63	-6.15	-5.77	-5.63	-8.35
β_p^H (eV)	-3.77765							
ζ_s^H (bohr ⁻¹)	1.08419	1.33	1.19	0.97	0.97	1.08	0.97	1.27
ζ_p^H (bohr ⁻¹)	0.88997							
α^H (Å ⁻¹)	3.16046	2.54	2.88	3.36	3.38	3.07	3.42	...
g_{ss}^H (eV)	12.65697	12.85	12.85	14.79	14.79	13.98	14.79	14.45
g_{sp}^H (eV)	11.34825							
g_{pp}^H (eV)	6.17416							
$g_{pp'}^H$ (eV)	10.0441							
h_{sp}^H (eV)	2.32560							
U_{ss}^O (eV)	-114.78169	-99.64	-97.83	-86.99	-87.41	-96.95	-86.96	-91.68
U_{pp}^O (eV)	-78.04828	-78.30	-78.26	-71.88	-72.18	-77.89	-71.93	-70.46
β_s^O (eV)	-31.51770	-32.69	-29.27	-45.20	-44.87	-29.85	-45.23	-65.64
β_p^O (eV)	-35.10436	-32.69	-29.27	-24.75	-24.60	-29.15	-24.79	-21.62
ζ_s^O (bohr ⁻¹)	3.19623	2.70	3.11	3.80	3.81	3.18	3.80	5.42
ζ_p^O (bohr ⁻¹)	3.11976	2.70	2.52	2.39	2.32	2.55	2.39	2.27
α^O (Å ⁻¹)	3.44202	3.16	4.46	3.22	3.23	4.17	3.39	...
g_{ss}^O (eV)	18.22143	15.42	15.42	15.76	15.76	14.00	15.76	11.30
g_{sp}^O (eV)	12.73220	14.52	14.52	10.62	10.62	14.96	10.62	15.81
g_{pp}^O (eV)	15.03924	14.48	14.48	13.65	13.65	14.15	13.65	13.62
$g_{pp'}^O$ (eV)	13.52768	12.98	12.98	12.41	12.41	12.70	12.41	10.33
h_{sp}^O (eV)	4.19786	3.94	3.94	0.59	0.59	3.93	0.59	5.01

specifically screened when the two atoms are in close proximity:

$$\begin{aligned} & \left\langle Hlm_l \left| -\frac{Z_B}{r_B} \right| Hlm'_l \right\rangle_{\text{PMO}} \\ &= \left\langle Hlm_l \left| -\frac{Z_B}{r_B} \right| Hlm'_l \right\rangle_{\text{MNDO}} [1 + P_2 \exp(-P_3 R_{HH}^2)] \end{aligned} \quad (5)$$

where P_2 is negative. Note that the standard MNDO approximation, $\langle Hlm_l | -Z_B/r_B | Hlm'_l \rangle_{\text{MNDO}}$, of this one-electron attraction integral is calculated from the two-electron integral $\langle BsOs0 | Hlm_l Hlm'_l \rangle$ (where we use the Mulliken convention for two-electron integrals), but the screening of eq 5 is applied to the one-electron integral (when B is a hydrogen and l is 1) and not to the two-electron integral.

- (3) For the homonuclear core–core repulsion integrals, we replace α^O and α^H by $\hat{\alpha}^O$ and $\hat{\alpha}^H$, respectively, whereas the heteronuclear core–core repulsions are computed from the atomic parameters α^O and α^H .

The parameters in the damped dispersion terms were retained from earlier work^{10–12,15} without change (in particular, they are the same as in MNDO-D and PM3-D from the Hillier group). MNDO parameters and the new PMO parameters (U_{pp}^H , β_p^H , ζ_p^H , g_{sp}^H , g_{pp}^H , $g_{pp'}^H$, h_{sp}^H , P_1 , P_2 , P_3 , $\hat{\alpha}^O$, and $\hat{\alpha}^H$) were adjusted by iterative optimizations using a genetic algorithm, in the presence of the dispersion term and the p orbitals on hydrogen atoms, to give a

Table 2. Additional PMO Parameters

parameter	value
P_1	0.15
P_2	-0.75
P_3	1.1 Å ⁻²
$\hat{\alpha}^O$	3.304 Å ⁻¹
$\hat{\alpha}^H$	2.466 Å ⁻¹
ζ_l^{HH}	1.280 bohr ⁻¹
ζ_l^{OO}	2.764 bohr ⁻¹

subjectively reasonable compromise in the accuracy of a lot of benchmark data. The final PMO values of the MNDO parameters, including those for p orbitals on H, are given in Tables 1 and 2. The parameters in Table 1 are those that may be compared to the parameters of other NDDO methods, and they are compared to MNDO,⁴ AM1,¹ PM3,² PDDG-PM3,²³ RM1,²⁴ PM3-D,¹⁰ and PM6.²⁵ Note that in Table 1, the parameters of PMO are given to the number of digits that define the parametrization, but for the other methods, the numbers are truncated to two places after the decimal point (calculations with the other methods employed the full number of digits, but the values for other methods in Table 1 are just for comparison, and they are rounded so that the essential features and trends are more apparent). The parameters in Table 2 are additional PMO parameters that do not appear in older methods.

The choice of functional forms in eqs 3a–5 was based on our experiences with parametrization. In particular, we tried several other functional forms and several other strategies for which terms to modify, and we selected the ones above because they are physically reasonable and they work well.

Table 3. Results

species	quantity	reference	source ^a	PMO	MNDO	AM1	PM3	PDDG/PM3	RM1	PM3-D	PM6	
H ₂ O	AE (kcal/mol)	232.2	27;28	233.0	225.0	223.0	217.2	245.4	221.6	246.5	218.0	
	IP (eV)	12.68	29;28,30	12.00	11.76	11.95	12.05	12.31	11.82	12.11	11.45	
	<i>r</i> (Å)	0.96	31	0.96	0.94	0.96	0.95	0.95	0.96	0.93	0.95	
	θ (deg)	104.5	31	104.6	106.8	103.5	107.7	105.4	103.4	108.9	107.5	
	α (Å ³)	1.45	32	1.24	0.43	0.50	0.50	0.49	0.51	0.46	0.40	
	q ^O	... ^b	...	-0.40	-0.33	-0.38	-0.36	-0.39	-0.37	-0.37	-0.62	
	q ^H	0.20	0.16	0.19	0.18	0.19	0.18	0.19	0.31	
	μ (Debye)											
	Mulliken	...	32	1.13	0.88	1.09	0.97	1.08	1.05	0.96	1.67	
	hybrid	1.06	0.90	0.77	0.77	0.75	0.82	0.78	0.40	
total	1.85	33	2.19	1.78	1.86	1.74	1.84	1.87	1.74	2.07		
(H ₂ O) ₂	BE (kcal/mol) ^c	5.0	34	4.7	1.0	5.5	3.5	3.7	2.8	6.5	4.9	
	<i>r</i> ₁₄ (Å)	0.97	35	0.97	0.94	n.q.c. ^d	0.96	0.97	n.q.c.	0.94	n.q.c.	
	<i>r</i> ₁₃ (Å)	0.96	35	0.96	0.94	n.q.c.	0.95	0.95	n.q.c.	0.93	n.q.c.	
	<i>r</i> ₂₄ (Å)	1.95	35	1.96	3.42	n.q.c.	1.81	1.71	n.q.c.	1.77	n.q.c.	
	<i>r</i> ₂₅ (Å)	0.96	35	0.97	0.94	n.q.c.	0.95	0.96	n.q.c.	0.93	n.q.c.	
	θ ₄₁₃ (deg)	104.5	35	103.9	106.8	n.q.c.	107.7	105.3	n.q.c.	108.8	n.q.c.	
	θ ₁₄₂ (deg)	172.9	35	173.4	114.0	n.q.c.	179.6	178.2	n.q.c.	172.1	n.q.c.	
	θ ₁₂₆ (deg)	110.4	35	112.0	130.4	n.q.c.	110.6	113.2	n.q.c.	113.3	n.q.c.	
(H ₂ O) ₂	μ (Debye)											
	Mulliken	...		1.76	1.80	2.34	1.63	1.96	2.20	1.78	1.49	
	hybrid	...		1.42	1.81	1.52	0.86	0.87	1.63	0.98	0.38	
total	2.65	36	3.11	3.60	3.86	2.49	2.83	3.83	2.76	1.87		
H ₂	AE (kcal/mol)	109.5	37	108.2	103.8	109.7	117.6	137.0	106.1	119.9	129.8	
	IP (eV)	15.55	38;39	13.60	14.72	14.19	15.53	15.85	14.33	15.56	14.17	
	<i>r</i> (Å)	0.74	37	0.70	0.66	0.68	0.70	0.68	0.70	0.69	0.76	
OH ⁻	AE (kcal/mol)	115.4	40,41;39	116.7	117.4	125.8	129.2	149.8	121.2	144.9	144.7	
	IP (eV)	1.83	42;39	2.18	0.27	0.65	0.89	1.16	0.55	0.92	2.05	
	<i>r</i> (Å)	0.96	43	0.88	0.94	0.94	0.94	0.94	0.94	0.92	0.87	
H ₃ O ⁺	<i>r</i> (Å)	0.98	44	1.02	0.96	1.00	0.98	0.98	0.99	0.95	1.04	
	θ (deg)	111.3	44	104.1	115.3	107.8	109.4	106.5	109.5	110.8	102.0	

^a When more than one source is cited, the source(s) before the semicolon is (are) for the value at 0 K, and the source(s) after the semicolon is (are) for the zero point energy used to convert it to a potential energy difference. ^b ... denotes not applicable. ^c BE = binding energy = 2(energy of the gas-phase water monomer optimized with the given method) - (energy of the gas-phase water dimer optimized with the given method). ^d The geometrical parameters of the water dimer are not given for AM1, RM1, and PM6 because the structure is not qualitatively correct (n.q.c.).

It is worthwhile to also mention another choice issue here, namely, the choice of p basis functions on hydrogen atoms rather than (for example) d basis functions on nonhydrogenic atoms, which may have seemed a more obvious choice. The introduction of d basis functions on nonhydrogenic atoms has been considered previously,²⁶ and these functions are known to improve the description of small rings and of some structures containing elements from the 3p block. Here our goal is an improved description of polarizability, even for compounds with atoms no heavier than the 2p block, and the choice of p basis functions on hydrogen is motivated by the results of part 1. Note that our goal is to achieve the best performance at a minimal increase in computational cost; this balance favors a set of p orbitals on hydrogenic atoms.

3. RESULTS

The computed results for molecules and for various water dimer configurations are listed in Tables 3–6, along with the reference data (i.e., the most accurate available data) for

comparison. The following abbreviations are used throughout: AE, atomization energy; IP, ionization potential; *r*, bond distance; θ , bond angle; α , polarizability; μ , dipole moment; and BE, bond energy. The reference data in Table 3 are based on previous work, usually experimental, but we needed to convert experimental energetic quantities (AE and BE) from enthalpy differences at 0 K to potential energy differences by subtracting zero point energies, based on experimental spectra or theoretical estimates. The third column of the table gives details of the sources^{27–44} of all reference data.

Table 3 shows that the polarizability of water is now much more accurate than that from previous NDDO methods. In fact, we can find parameters that make it perfect, while still yielding reasonable results for other data, but our final parameters are a compromise in which we accept a 14% error in water's polarizability in order to retain good accuracy for other quantities.

Table 3 also shows remarkably good accuracy for the atomization energy (AE) of water, the bond energies of key diatomics, and the ionization potential (IP) and the geometry of water and key diatomics. Fitting simultaneously the dipole moment and

Table 4. Potential Curves^a for Water Dimer in Hydrogen Binding Orientation (kcal/mol)

$d(\text{H}-\text{O})$ (Å)	PMO	AM1	MNDO	PM3	RM1	PM6	PM3-D	PDDG-PM3	M06-2X/MG3S	CCSD(T)-F12b/aug-cc-pVTZ
0.948	83.1	88.9	122.1	61.9	72.3	74.9	39.6	61.2	82.0	83.6
1.248	16.6	26.2	47.1	22.2	13.6	15.2	11.3	17.7	15.5	16.8
1.448	2.6	11.8	27.2	9.3	2.9	1.6	2.5	4.7	1.7	2.4
1.648	-2.7	3.0	15.9	-1.3	-0.5	-2.9	-5.6	-3.3	-3.7	-3.5
1.948	-4.6	-2.8	6.7	-2.8	-0.9	-3.9	-5.1	-1.1	-5.5	-5.1
2.448	-3.7	-2.5	0.8	-1.7	-1.8	-2.8	-2.5	-1.9	-4.0	-3.8
2.948	-2.5	-1.3	-0.5	-1.1	-1.2	-1.7	-1.5	-1.3	-2.5	-2.3
3.948	-1.1	-0.6	-0.5	-0.5	-0.6	-0.8	-0.6	-0.6	-1.0	-0.9
4.948	-0.6	-0.3	-0.3	-0.3	-0.3	-0.5	-0.3	-0.3	-0.5	-0.5
5.948	-0.3	-0.2	-0.2	-0.2	-0.2	-0.3	-0.2	-0.2	-0.3	-0.3
6.948	-0.2	-0.1	-0.1	-0.1	-0.1	-0.2	-0.1	-0.1	-0.2	-0.2

^aThe energies shown are relative to the energy of the dimer when $d(\text{H}-\text{O}) = 99$ Å.

Table 5. Potential Curves^a for Water Dimer in Repulsive Overlay Orientation (kcal/mol)

$d(\text{O}-\text{O})$ (Å)	PMO	AM1	MNDO	PM3	RM1	PM6	PM3-D	PDDG-PM3	M06-2X/MG3S	CCSD(T)-F12b/aug-cc-pVTZ
2.0	40.5	12.9	44.4	13.7	23.2	16.3	5.9	20.4	43.5	44.5
2.5	10.0	1.2	11.3	4.0	2.7	3.9	1.2	2.9	9.1	10.1
3.0	3.5	1.5	3.6	1.7	1.6	2.5	0.6	1.6	2.9	3.1
4.0	1.3	0.8	0.8	0.7	0.8	0.9	0.5	0.7	0.9	0.9
5.0	0.7	0.4	0.4	0.4	0.4	0.5	0.3	0.4	0.5	0.5
6.0	0.4	0.2	0.2	0.2	0.2	0.3	0.2	0.2	0.3	0.3

^aThe energies shown are relative to the energy of the dimer when $d(\text{O}-\text{O}) = 99$ Å.

Table 6. Interaction Energies (kcal/mol) at Water Dimer Stationary Points^a

structure	PMO	AM1	MNDO	PM3	RM1	PM6	PM3-D	PDDG-PM3	reference ^b
nonplanar open C_s	-4.7	-2.8	7.5	-2.4	-0.9	-3.7	-2.3	-1.0	-5.0
open C_1	-4.1	-2.2	7.3	-1.4	-0.2	-2.9	-1.2	0.3	-4.5
planar open C_s	-4.1	-2.4	6.7	-1.4	-0.5	-3.0	-1.1	0.4	-4.4
cyclic C_i	-2.7	-3.8	7.6	-0.2	-0.2	-3.4	0.2	0.8	-4.3
cyclic C_2	-2.3	-3.4	7.8	0.3	0.2	-2.9	0.7	1.2	-4.0
cyclic C_{2h}	-2.3	-3.5	7.6	0.4	0.0	-3.1	0.7	1.3	-4.0
triply H bonded C_s	-2.7	-3.8	3.7	-1.2	-2.6	-2.8	-0.6	-1.1	-3.2
doubly bifurcated C_{2h}	-1.3	-1.6	1.9	-0.6	-1.3	-1.0	0.5	-1.5	-1.4
nonplanar bifurcated C_{2v}	-3.0	-3.7	2.8	-1.2	-2.8	-3.0	-0.3	-2.3	-3.2
planar bifurcated C_{2v}	-2.6	-2.9	1.6	-1.1	-2.5	-2.3	0.0	-2.4	-2.3
mean unsigned deviation	0.7	1.0	9.1	2.7	2.6	0.8	3.3	3.2	

^aAll interaction energies in this table are with respect to two separated monomers at the monomer geometry of ref 35. ^bCoupled cluster value from ref 35.

molecular polarizability of water is problematic and requires a compromise in consideration of various aspects of applications in the gas phase and in condensed phases. On one hand, the total dipole moment is calculated from the wave function, which is a sum of two contributions in the NDDO approximation: one from the partial charges formally equivalent to those by Mulliken analysis and one from the hybridization terms in the dipole matrix elements. On the other hand, the wave functions of semiempirical models that employ a minimal basis set tend to severely underestimate the charge separations of polar bonds, resulting in Mulliken population charges or electrostatic-potential-fitted charges that are too small for use in condensed-phase simulations. From the latter perspective, it is not critical to enforce a perfect agreement with the experimental dipole moment of an isolated water molecule in the gas phase; in fact, it is

desirable to have a somewhat enhanced molecular dipole moment in view of the limitations of a minimal basis set for describing polar bonds. The compromise we chose is to simply require that the experimental dipole moment lies between the Mulliken and total values.

The final data to be discussed in Table 3 are those for the water dimer. Several previous NDDO methods give qualitatively incorrect geometries for the water dimer. The water dimer structures for all of the NDDO methods are shown in Figure 1, and these structures may be compared to the best estimate³⁵ of the water dimer structure, which is shown in Figure 2. PMO not only gives the qualitatively correct structure but gives reasonably accurate geometrical parameters, as shown in Table 3. Furthermore, the reference value for the dipole moment of the water dimer is between the Mulliken and total values calculated by PMO.

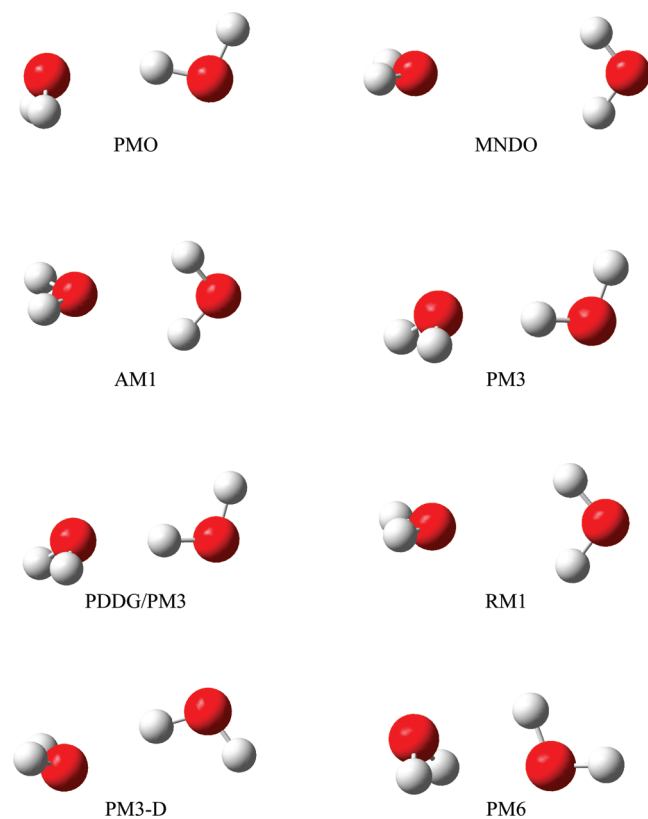


Figure 1. Equilibrium water dimer structure predicted by various semiempirical electronic structure methods.

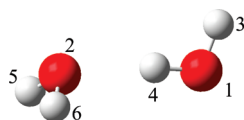


Figure 2. Best estimate of the structure of the minimum-energy water dimer from ref 35.

A key issue in parametrizing semiempirical methods is to achieve a realistic combination of noncovalent attractive interactions and exchange repulsion. This has led to a variety of special terms being added to the core–core repulsion, including Gaussian modification terms.^{1,2,23} The performance of PMO to predict the interaction energy between two water molecules over a wide range of separation distances is discussed next.

We consider two relative orientations of the water molecules, one being the hydrogen bonded orientation and the other, shown in Figure 3, being a structure that we will label the repulsive overlay. For the hydrogen bonded orientation, we start with the optimum dimer structure of Tschumper et al., and we pull the monomers apart along the O–H hydrogen bond direction. For the repulsive overlay, we make the planes of the two monomers parallel, with one water molecule precisely aligned with the other one. In the first case, the interaction energy is determined as a function of the O–H hydrogen bond distance, $d(\text{O–H})$, and in the second, it is calculated as a function of the O–O distance, $d(\text{O–O})$. In both cases, we calculated the potential energy curve, relative to the separated monomers, by two high-level methods. Note that the monomers are rigid in both structures. For the hydrogen bonded structures, they are rigid at the geometries they

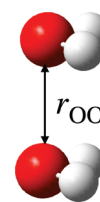


Figure 3. “Repulsive overlay” water dimer.

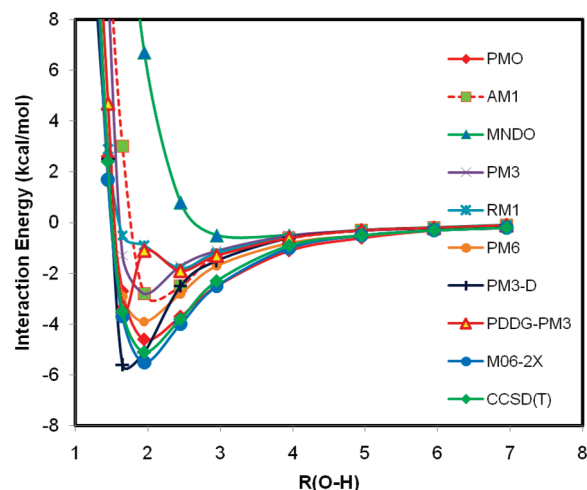


Figure 4. Potential energy curves (kcal/mol) for the hydrogen-bonding configuration of a water dimer computed using various semiempirical, ab initio wave function, and density functional methods, as labeled in the figure. Distances are shown in angstroms.

have in the dimer structure of Tschumper et al., and for the repulsive overlay they are rigid at the experimental geometry⁴⁵ of the gas-phase water monomer. We use the same geometries for the MNDO calculations and the high-level calculations. The high-level calculations are carried out by the M06-2X density functional⁴⁶ with the MG3S⁴⁷ basis and by the high-level CCSD(T)-F12b⁴⁸ ab initio wave function method with the aug-cc-pVTZ⁴⁹ basis set. The latter should be close to the complete configuration interaction limit.

Table 4 and Figure 4 show that only PMO and PM6 give reasonably accurate potential curves for the hydrogen bonding orientation, although the binding energy from PM6 is 1 kcal/mol too weak in comparison with CCSD(T) results. The other methods are not attractive enough, as seen by their prediction of highly repulsive interactions at hydrogen-bonding distances, a shortfall of the original MNDO method that led to the introduction of Gaussian terms in the core–core potential in the highly successful AM1 model. PM3-D was developed by adjusting some of the original PM3 parameters after damped dispersion terms are introduced, which yields an excellent binding energy, but the hydrogen-bond distance at the minimum-energy geometry is too short, and the potential energy curve dies off too quickly at medium distances. The recently introduced RM1 model produced a minimum that is too far away for a good single hydrogen-bonded complex and a shoulder with a much higher energy at a distance shorter than the best estimate of the minimum. This is reflected in the minimum energy configuration obtained with RM1 (Figure 1), showing a bifurcated complex—a deficiency that is also well-known for a water dimer in the MNDO and AM1 models. Table 5 shows that only PMO and MNDO give

reasonable potential energy curves in the repulsive overlap orientation, with the other curves not being repulsive enough. Thus, the present parametrization succeeds in representing these opposing types of interactions well, over a wide range of distances.

Next, we consider a broad set of water dimer geometries. In particular, Table 6 gives results at the geometries of all 10 stationary points on the water dimer surface that were characterized by Tschumper et al.³⁵ at the coupled cluster level, where only one structure is a minimum and the others consist of three first-order saddle points and six higher-order saddle points

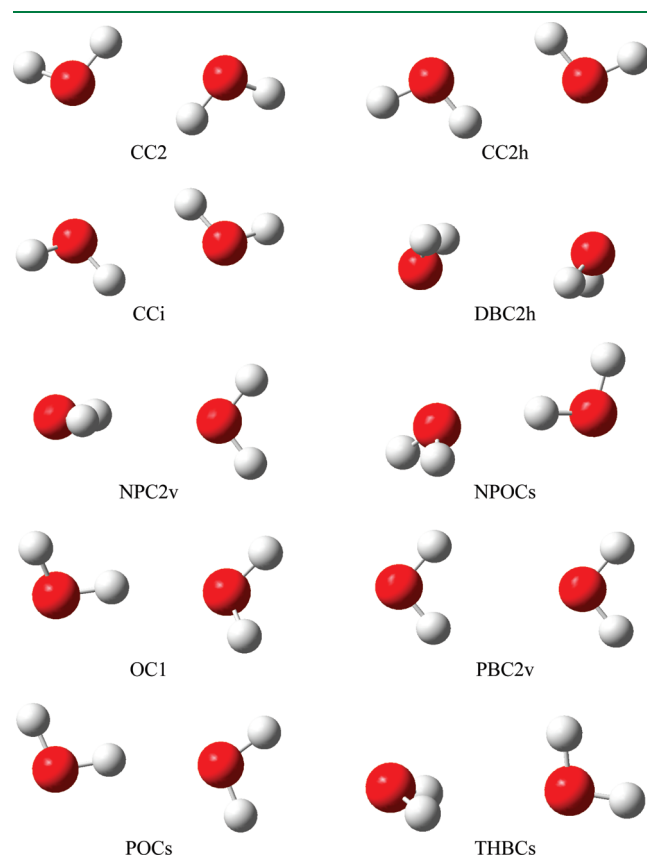


Figure 5. Stationary points on the water dimer potential energy surface from ref 35.

(hilltops); the names of the structures are those given to them in ref 35, and the structures are illustrated in Figure 5. Our comparisons are carried out at these geometries without any reoptimization; the goal is not to test what set of stationary structures is predicted by the methods but rather to test whether the potential energy surface is well reproduced over a range of geometries at various locations on the high-dimensional potential energy surface. The table shows that PMO gives a qualitatively correct description of the various interaction energies, and it is in better agreement with the accurate results than any previous NDDO method, although AM1 and PM6 do almost as well.

Tables 7 and 8 present results for larger water clusters, up to the octamer. These larger clusters are a severe test of the ability of the PMO model to include nonadditive polarization effects such as those that occur when a monomer interacts with other monomers in more than one direction. Table 7 shows the energies computed with each semiempirical model using the structures from the work of Bryantsev et al.,⁵⁰ who used density functional theory (DFT) geometries and relative energies obtained by high-level (CCSD(T)) ab initio wave function theory.

Whereas Table 7 is a comparison of interaction energies using the set of fixed structures from ref 50, Table 8 provides a different kind of test. For Table 8, the results were obtained with the fully optimized structures for these clusters at each semiempirical level; the structures of ref 50 were used as the starting points, and the nearest lower-energy stationary points were sought for each semiempirical model using the default geometry optimizer in the MOPAC⁵¹ software program, which is based on a modified Broyden–Fletcher–Goldfarb–Shanno (BFGS) method.⁵² (A configuration was considered optimized when its gradient norm fell below $0.5 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$.) The PMO optimized structures obtained by this procedure are shown in Figure 6, whereas those from other models are supplied in the Supporting Information. Only the PMO calculations yield cluster structures that closely resemble those of the DFT results of Bryantsev et al.⁵⁰ Furthermore, for PMO, the agreement is very good for all cluster configurations. The data in Table 7 are quite revealing of the deficiencies exhibited in each semiempirical model. As expected on the basis of results for the dimer complex (Table 4), MNDO gives strongly repulsive interactions for these clusters at hydrogen-bonding geometries. In fact, at these fixed hydrogen-bonding configurations, none of the previous

Table 7. Interaction Energies (kcal/mol) of Water Clusters from Single-Point Calculations Using the Fixed Geometries from Ref 50^a

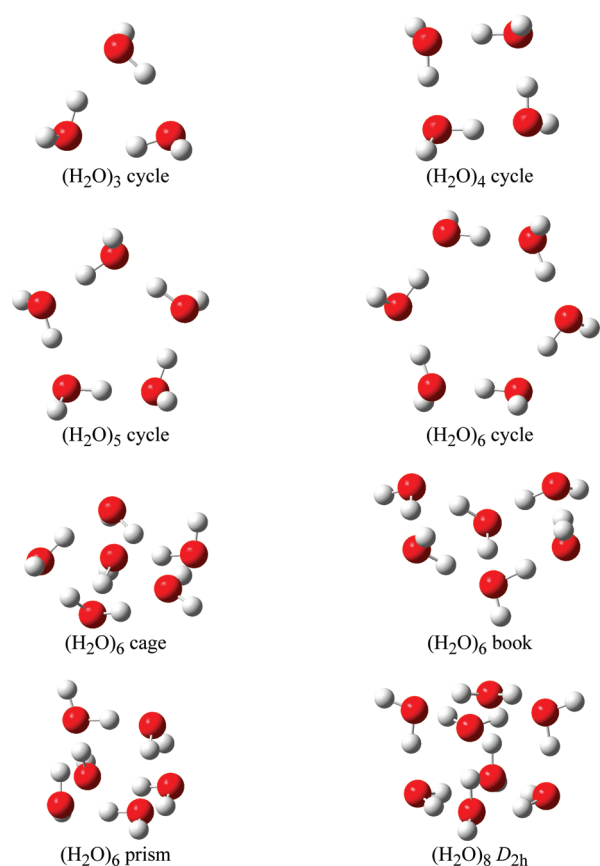
structure	PMO	AM1	MNDO	PM3	RM1	PM6	PM3-D	PDDG-PM3	reference
(H ₂ O) ₂	-4.6	-2.9	7.1	-2.5	-0.9	-3.8	-4.3	-0.9	-5.0
(H ₂ O) ₃ cycle	-8.3	-2.3	40.6	-9.2	-1.4	-11.8	-20.9	-7.5	-15.8
(H ₂ O) ₄ cycle	-25.1	-5.8	45.0	-17.8	-6.7	-20.7	-28.5	-17.1	-27.4
(H ₂ O) ₅ cycle	-28.1	2.2	70.2	-18.6	-9.4	-24.5	-40.0	-25.1	-35.9
(H ₂ O) ₆ cycle	-41.3	-4.9	68.6	-27.5	-14.4	-32.8	-44.6	-30.6	-44.3
(H ₂ O) ₆ cage	-42.2	-16.0	78.0	-24.5	-7.0	-33.9	-44.2	-17.8	-46.0
(H ₂ O) ₆ book	-42.2	-10.1	74.8	-26.8	-10.3	-33.5	-45.2	-24.6	-45.8
(H ₂ O) ₆ prism	-42.3	-21.1	76.5	-21.9	-7.4	-34.9	-41.8	-14.4	-45.3
(H ₂ O) ₈ D _{2h}	-69.1	-22.6	121.4	-39.5	-12.4	-53.9	-70.6	-30.3	-72.6
mean unsigned deviation	3.9	28.3	102.3	16.6	29.8	9.8	2.1	18.9	

^a Structures and reference energies from ref 50. The interaction energy is calculated relative to the infinitely separated gas-phase monomers with geometries frozen in the monomer geometry given in ref 50.

Table 8. Interaction Energies (kcal/mol) of Water Clusters Computed Using the Optimized Structures at Each Theoretical Model^a

Structure	PMO	AM1	MNDO	PM3	RM1	PM6	PM3-D	PDDG-PM3	reference
(H ₂ O) ₂	-4.7	-5.5	-1.0	-3.5	-2.8	-4.9	-6.5	-3.7	-5.0
(H ₂ O) ₃ cycle	-13.4	-15.3	-1.6	-10.1	-4.0	-13.3	-19.8	-10.2	-15.8
(H ₂ O) ₄ cycle	-26.0	-22.1	-3.1	-18.3	-10.0	-22.3	-31.8	-21.1	-27.4
(H ₂ O) ₅ cycle	-34.7	-31.8	-4.7	-23.8	-14.5	-29.4	-40.8	-28.0	-35.9
(H ₂ O) ₆ cycle	-42.6	-32.8	-5.1	-29.2	-18.3	-36.0	-49.1	-34.5	-44.3
(H ₂ O) ₆ cage	-43.9	-30.2	-4.8	-30.2	-16.1	-37.5	-60.9	-34.7	-46.0
(H ₂ O) ₆ book	-43.6	-38.9	-5.0	-30.0	-17.7	-35.1	-54.6	-34.9	-45.8
(H ₂ O) ₆ prism	-43.9	-39.4	-4.3	-29.7	-16.4	-39.3	-62.5	-35.0	-45.3
(H ₂ O) ₈ D _{2h}	-72.4	-56.5	-6.9	-48.6	-24.0	-58.1	-93.8	-56.6	-72.6
Mean unsigned deviation	1.5	7.4	33.5	12.8	23.8	6.9	9.1	8.8	

^aThe starting geometries in all structural optimizations are taken from the DFT optimized geometries in ref 50. Reference energies are taken from ref 50. The interaction energy is calculated relative to the infinitely separated gas-phase monomers with geometries optimized with the given method.

**Figure 6.** Optimized water clusters using the PMO method.

semiempirical models are attractive enough in comparison with CCSD(T) results, except PMO and PM3-D, in which dispersion energies are explicitly modeled. The interaction energies for the optimized structures do not show noticeable improvement after the structures are fully relaxed in each previous semiempirical model; as it turns out, AM1 and PM6 perform the best for the larger clusters, whereas the binding energies from PM3-D become too large for the larger water clusters. Table 8 shows that the mean unsigned error of the present PMO method is only 1.5 kcal/mol for these water clusters. The success of the PMO model for the cluster energetics at consistently optimized geometries is a crowning achievement for the method.

Table 9. Dispersion Contributions to the Energies of Water Clusters^a

structure	at geometries of Table 7				at geometries of Table 8			
	PMOxD	D	PMO	%D	PMOxD	D	PMO	%D
(H ₂ O) ₂	-3.8	-0.8	-4.6	17	-3.9	-0.8	-4.7	17
(H ₂ O) ₃ cycle	-5.9	-2.4	-8.3	29	-11.0	-2.4	-13.4	18
(H ₂ O) ₄ cycle	-21.3	-3.8	-25.1	15	-21.9	-4.1	-26.0	16
(H ₂ O) ₅ cycle	-23.2	-4.9	-28.1	17	-29.6	-5.1	-34.7	15
(H ₂ O) ₆ cycle	-35.6	-5.7	-41.3	13	-36.6	-6.0	-42.6	14
(H ₂ O) ₆ cage	-33.6	-8.6	-42.2	20	-35.0	-8.9	-43.9	20
(H ₂ O) ₆ book	-35.0	-7.2	-42.2	17	-36.3	-7.3	-43.6	17
(H ₂ O) ₆ prism	-33.2	-9.1	-42.3	22	-34.8	-9.1	-43.9	21
(H ₂ O) ₈ D _{2h}	-55.7	-13.4	-69.1	19	-57.9	-14.5	-72.4	20
mean percentage	19	18

^aEnergies are in kcal/mol, and %D is defined as (D/PMO) × 100%.

A question of interest in parametrization is how much the damped dispersion term contributes to the results. This is illustrated for the water clusters in Table 9. In this table, PMOxD is the result obtained excluding the damped dispersion term. D denotes the contribution of the damped dispersion term, and PMO is the total. Results are shown both for the accurate geometries of Table 7 and for the consistently optimized geometries of Table 8. We see that the dispersion contribution ranges from 13 to 22% of the total binding energy, with the percentage not depending strongly on cluster size. The average, shown in the last row, is almost 20%. We may compare the present result for the damped dispersion energy at the equilibrium geometry of the dimer to the percentage estimated by the ab initio second-order symmetry-adapted perturbation theory (SAPT2), which yields a damped dispersion contribution of -2.49 kcal/mol, or 46% of the total interaction energy of the cluster (-5.4 kcal/mol) computed at that level of approximation.⁵³ Note that the damped dispersion component is not uniquely defined; the quantity called dispersion in SAPT2 includes all correlation effects on the direct induction energy⁵³ (and hence it may also be called the correlation contribution to the direct induction/dispersion). Su and Li⁵⁴ use different definitions; they label the sum of all four correlation components of SAPT2 (direct static, direct induction/dispersion, exchange static, and exchange remainder) as dispersion; applying the

Table 10. Proton Transfer Energies (kcal/mol) for $A + H_3O^+ \rightarrow AH^+ + H_2O^a$

A	PMO	MNDO	AM1	PM3	PDDG/PM3	RM1	PM3-D	PM6	reference	source
H ₂	27.7	63.8	26.4	16.1	4.9	27.8	25.4	-7.3	66.7	51
OH ⁻	-220.0	-250.3	-247.8	-248.4	-248.3	-246.6	-249.8	-216.7	-223.1	27, 41, 42, 52
OH	17.1	14.6	12.6	8.9	8.7	13.2	8.3	0.1	26.0	41
H ₂ O ₂	7.6	20.5	11.7	3.2	9.6	15.8	8.5	0.5	9.7	41, 42, 53
HO ₂	13.8	17.6	10.2	0.5	2.0	10.8	3.6	-7.1	7.8	51
O ₃	20.3	109.5	108.5	100.5	111.0	106.1	97.4	-8.9	17.4	51
MUD ^b	10.3	25.7	29.0	31.7	34.0	28.8	28.5	26.1		

^a For all NDDO calculations in this table, the geometry of each reactant and product was optimized at the level under consideration. ^b Mean unsigned deviation from reference values, which are based on the sources indicated in the last column.

Su–Li definition⁵⁴ to the Rybak et al.⁵³ calculation would yield a “dispersion” contribution of -1.9 kcal/mol or 35%. Given the nonuniqueness of the damped dispersion contribution, the present results in Table 8 are physically reasonable.

The relative proton affinities (excluding, as in all other comparisons in this article, the zero point and thermal vibrational energy) are examined by considering the energy change of proton transfer reactions from an oxonium ion to different bases. Since these reactions involve charge migrations, they present another challenge for methods to adequately treat polarization effects. Table 10 compares the results of NDDO calculations to the reference data, which are obtained by combining data for A and AH⁺ (in the notation of the table heading) from a number of sources^{27,41,42,55–57} with the proton affinity of water⁵⁸ and removing vibrational contributions. The table shows that PMO yields better agreement with the reference results than do previous NDDO models.

4. DISCUSSION

As mentioned in section 2, the parametrization reported here represents a compromise. For example, we could find parameters that give the precisely correct polarizability and dipole moment of water; however, it would yield somewhat less accurate interaction energy curves for water dimers and binding energies of water cluster than those in Tables 3–5. We also compromised on fitting the dipole moment of water, simply requiring the accurate value (1.85 D) to lie between the value (1.13 D) calculated from the Mulliken charges and the value (2.19 D) calculated when including the so-called hybrid terms (that is, the atomic dipole terms). Another compromise is that we could obtain more accurate results for water dimers and clusters, but at the expense of lowering the polarizability of water and allowing the dipole moment of water to deviate somewhat from the experimental value. Our final compromise is to require that the polarizability of water is accurate to within 15%.

The inclusion of polarization in force fields used for molecular simulation makes them more realistic, but including polarization in a molecular mechanics framework often involves approximations of uncertain physicality such as the introduction of artificial charge centers, dual thermostatting in shell-type models, and the need to partition the polarizability into atomistic contributions.^{59–64} The quantum mechanical framework presented here avoids such devices and represents the polarizability naturally in an SCF framework. Polarization is an important consideration in parametrizing model chemistries. In a liquid, the dielectric screening of electrostatics on average increases the magnitudes of partial atomic charges and dipole moments of neutral molecules. Thus, for a nonpolarizable model, the effective charges are

typically parametrized to yield molecular dipole moments about 15% to 20% greater than those in the gas phase; ideally this value would be determined by carrying out simulations on liquids. With polarizable force fields, there is a much greater chance of incorporating polarization effects in response to the instantaneous fluctuations of surrounding solvent and in response to changes in the local electrical field, and thus there is a greater chance that a model can be reasonable both for molecules and small clusters, on the one hand, and for liquids, on the other. In fact, the success of the present model for water clusters ranging from dimers to octamers is already encouraging. Future tests on liquid water would be very interesting.

The parameters presented here are illustrative of the performance with the present theoretical model in its simplest form. It is anticipated that further refinement in parametrization and improved functional form will be found in future work. For example, the two-center terms can be optimized specifically for each pair as a function of the internuclear distance rather than using atomic parameters, or different functional forms may be adopted for different pairs; for example, one can use the pairwise core–core repulsion form suggested by Voityuk and Röscher.⁶⁵ Clearly, the next step is to extend the current model to a much larger data set and to a broader range of functional groups. The dispersion terms used in the present PMO model were taken directly from previous work; other functional forms^{66–68} and parameters may be fully optimized for complex systems. For condensed-phase simulations, such as liquid water, it is critical to evaluate and further optimize the noncovalent attractive and repulsive potentials in the framework of the X-Pol method. As a model for the development of a next-generation force field for macromolecular simulations, the accuracy may be further improved by letting parameters depend on atom type and hybridization state—such a parametrization would be used only for nonreactive systems or nonreactive subsystems.

The p orbital exponential parameter for hydrogen used here is $\zeta = 0.88997 a_0^{-1}$. One may convert this to an equivalent Gaussian exponential parameter α by the STO-1G prescription, which yields²¹

$$\alpha = 0.175967\zeta^2$$

One then obtains $\alpha = 0.13937 a_0^{-2}$, which compares well to the value $\alpha = 0.141 a_0^{-2}$ used for the diffuse p orbital in the STO-3G(P) basis set of paper 1 even though no such correspondence was used in optimizing the parameters of PMOv1.

5. CONCLUDING REMARKS

We have developed a new parametrized semiempirical molecular orbital model that includes polarization effects much more

realistically than previous parametrizations, and we illustrated it for calculations on systems composed of oxygen and hydrogen atoms, such as water clusters and various protonated hydrogen–oxygen compounds.

The difficulty in modeling hydrogen bonding interactions using semiempirical models is well-known, and it has been a continuing challenge since the original MNDO and subsequent AM1 and PM3 methods. Despite many efforts in parametrization in the past 40 years, even with the very recent and carefully parametrized PM6, the ability to describe hydrogen bonding interactions remains the weakest feature in semiempirical methods and the most difficult to improve. To extend semiempirical methods to model biological systems as a quantal force field, this is the key issue that must be solved. The present papers demonstrated that there is a good theoretical foundation (paper 1) and an excellent practical opportunity (paper 2) to make this happen. We note that no other semiempirical methods predict the water cluster geometries and energies even remotely close to the ab initio results; the present PMO method can yield good results for both the geometries and the energies of these clusters.

The new method is available in a version of MOPAC distributed free of charge on the Internet.⁵¹

■ ASSOCIATED CONTENT

S Supporting Information. Additional structures as in Figure 6 but as obtained with each of the NDDO methods. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: truhlar@umn.edu (D.G.T.), gao@jialigao.org (J.G.).

■ ACKNOWLEDGMENT

This work was supported in part by the National Institutes of Health (grant no. RC1-GM091445) and the National Science Foundation (grant no. CHE09-56776).

■ REFERENCES

- (1) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
- (2) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 209.
- (3) Pople, J. A.; Santry, D. P.; Segal, G. A. *J. Chem. Phys.* **1965**, *43*, S129.
- (4) Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4899.
- (5) Gao, J. *J. Phys. Chem. B* **1997**, *101*, 657. Xie, W.; Gao, J. *J. Chem. Theory Comput.* **2007**, *3*, 1890.
- (6) Xie, W.; Song, L.; Truhlar, D. G.; Gao, J. *J. Chem. Phys.* **2008**, *128*, 234108. Xie, W.; Orozco, M.; Truhlar, D. G.; Gao, J. *J. Chem. Theory Comput.* **2009**, *5*, 459.
- (7) Fiedler, L.; Gao, J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2011**, DOI: 10.1021/ct1006373.
- (8) Parkinson, W. A.; Zerner, M. C. *J. Chem. Phys.* **1991**, *94*, 478.
- (9) Bredow, T.; Jug, K. *Theor. Chem. Acc.* **2005**, *113*, 1.
- (10) McNamara, J. P.; Hillier, I. H. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2362.
- (11) Morgado, C. A.; McNamara, J. P.; Hillier, I. H.; Burton, N. A.; Vincent, M. A. *J. Chem. Theory Comput.* **2007**, *3*, 1656.
- (12) McNamara, J. P.; Sharma, R.; Vincent, M. A.; Hillier, I. H.; Morgado, C. A. *Phys. Chem. Chem. Phys.* **2008**, *10*, 128.
- (13) Roothaan, C. C. *J. Rev. Mod. Phys.* **1951**, *23*, 69.
- (14) Mitoraj, M. P.; Michalak, A.; Ziegler, T. *J. Chem. Theory Comput.* **2009**, *5*, 962.
- (15) Jeziorski, B.; Moszynski, R.; Szalewicz, K. *Chem. Rev.* **1994**, *94*, 1887.
- (16) Bickelhaupt, F. M.; Baerends, E. J. *Rev. Comp. Chem.* **1999**, *15*, 1.
- (17) Mo, Y.; Gao, J.; Peyerimhoff, S. D. *J. Chem. Phys.* **2000**, *112*, 5530.
- (18) Wu, Q.; Yang, W. *J. Chem. Phys.* **2002**, *116*, 515.
- (19) Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1463.
- (20) Dewar, M. J. S.; Thiel, W. *Theor. Chim. Acta* **1977**, *46*, 89.
- (21) Stewart, R. J. *J. Chem. Phys.* **1970**, *52*, 431.
- (22) Dewar, M. J. S.; Low, D. H. *J. Am. Chem. Soc.* **1972**, *94*, 5296.
- (23) Repasky, M. P.; Chandrasekhar, J.; Jorgensen, W. L. *J. Comput. Chem.* **2002**, *23*, 1601.
- (24) Rocha, G. B.; Freire, R. O.; Simas, A. M.; Stewart, J. J. P. *J. Comput. Chem.* **2006**, *27*, 1101.
- (25) Stewart, J. J. P. *J. Mol. Model.* **2007**, *13*, 1173.
- (26) Thiel, W.; Voityuk, A. A. *J. Phys. Chem.* **1996**, *100*, 616. Hawkins, G. D.; Cramer, D. J.; Truhlar, D. G. *J. Phys. Chem. B* **1998**, *102*, 3257. Hutter, M. C.; Reimers, J. R.; Hush, N. S. *J. Phys. Chem. B* **1998**, *102*, 8080. Jomoto, L. J.; Nakajima, T. *THEOCHEM* **2002**, *577*, 143. Lopez, X.; York, D. M. *Theor. Chem. Acc.* **2003**, *109*, 149. Dybala-Defratyka, A.; Paneth, P.; Pu, J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 2475. Kwiecień, R. A.; Rostkowski, M.; Dybala-Defratyka, A.; Paneth, P. *J. Inorg. Biochem.* **2004**, *98*, 1078. Nam, K.; Cui, Q.; Gao, J.; York, D. M. *J. Chem. Theory Comput.* **2007**, *3*, 486. Tejero, I.; González-Lafont, J.; Lluch, J. M. *J. Comput. Chem.* **2007**, *28*, 997. Sorkin, A.; Truhlar, D. G.; Amin, E. A. *J. Chem. Theory Comput.* **2009**, *5*, 1254.
- (27) Pople, J. A.; Head-Gordon, M.; Fox, D. J.; Raghavachari, K.; Curtiss, L. A. *J. Chem. Phys.* **1989**, *90*, 5622.
- (28) Martin, J. M. L. *J. Chem. Phys.* **1992**, *97*, 5012.
- (29) Page, R. H.; Larkin, R. J.; Shen, Y. R.; Lee, Y. T. *J. Chem. Phys.* **1988**, *88*, 2249.
- (30) The H₃O⁺ zero point energy was calculated using the MC-QCISD/3 method: Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2003**, *107*, 3898. with a scale factor of 0.994, as recommended: Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 6908.
- (31) *CRC Handbook of Chemistry and Physics, 2010–2011*, 91st ed.; Haynes, W. M., Ed.; CRC Press: New York, 2010; p 9–24.
- (32) *Landolt-Bornstein Zahlenwerte und Funktionen*; Springer: Berlin, 1962; Vol 6, Aufl. Band II/8 S.6, p 871.
- (33) Dyke, T. R.; Muentzer, J. S. *J. Chem. Phys.* **1973**, *59*, 3125.
- (34) Dahlke, E. E.; Orthmeyer, M. A.; Truhlar, D. G. *J. Phys. Chem. B* **2008**, *112*, 2372.
- (35) Tschumper, G. S.; Leininger, M. L.; Hoffman, B. C.; Valeev, E. F.; Schaefer, H. F., III; Quack, M. *J. Chem. Phys.* **2002**, *116*, 690.
- (36) Bouteiller, Y.; Desfrancois, C.; Abdoul-Carime, H.; Schermann, J. P. *J. Chem. Phys.* **1996**, *105*, 6420.
- (37) Kolos, W.; Wolniewicz, L. *J. Chem. Phys.* **1968**, *49*, 404.
- (38) Shiner, D.; Gilligan, J. M.; Cook, B. M.; Lichten, W. *Phys. Rev. A* **1993**, *47*, 4042.
- (39) Herzberg, G. *Molecular Spectra and Molecular Structure. I. Diatomic Molecules*; Prentice-Hall, Inc.: New York, 1939; pp 487–489.
- (40) Neumark, D. M.; Lykke, K. R.; Anderson, T.; Lineberger, W. C. *Phys. Rev. A* **1985**, *32*, 1890. *Landolt-Bornstein Zahlenwerte und Funktionen*; Springer: Berlin, 1962; Vol 6, Aufl. Band II/8 S.6, p 871.
- (41) Ruscic, B.; Wagner, A. F.; Harding, L. B.; Asher, R. L.; Feller, D.; Dixon, D. A.; Peterson, K. A.; Song, Y.; Qian, X.; Cheuk-Yiu, N.; Liu, J.; Chen, W.; Schwenke, D. W. *J. Phys. Chem. A* **2002**, *106*, 2727.
- (42) Celotta, R. J.; Bennett, R. A.; Hall, J. L. *J. Chem. Phys.* **1974**, *60*, 1740.
- (43) Rosenbaum, N. H.; Owrutsky, J. C.; Tack, L. M.; Saykally, R. J. *J. Chem. Phys.* **1986**, *84*, 5308.
- (44) Sears, T. J.; Bunker, P. R.; Davies, P. B.; Johnson, S. A.; Spirko, V. *J. Chem. Phys.* **1985**, *83*, 2676.

- (45) Polyansky, O. L.; Jensen, P.; Tennyson, J. *J. Chem. Phys.* **1994**, *101*, 7651.
- (46) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215.
- (47) Lynch, B. J.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2003**, *107*, 1384.
- (48) Knizia, G.; Adler, T. B.; Werner, H.-J. *J. Chem. Phys.* **2009**, *130*, 054104.
- (49) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6769.
- (50) Bryantsev, V. S.; Diallo, M. S.; van Duin, A. C. T.; Goddard, W. A., III. *J. Chem. Theory. Comput.* **2009**, *5*, 1016.
- (51) Stewart, J. J. P.; Fiedler, L. J.; Zhang, P.; Zheng, J.; Rossi, I.; Hu, W.-P.; Lynch, G. C.; Liu, Y.-P.; Chuang, Y.-Y.; Pu, J.; Li, J.; Cramer, C. J.; Fast, P. L.; Gao, J.; Truhlar, D. G. *MOPAC*, version 5.017mn (2010); University of Minnesota: Minneapolis and Saint Paul, MN. This program is available at <http://comp.chem.umn.edu/mopac> (accessed Jan 2011).
- (52) Stewart, J. J. P. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1.
- (53) Rybak, S.; Jeziorski, B.; Szalewicz, K. *J. Chem. Phys.* **1991**, *95*, 6576.
- (54) Su, P.; Li, H. *J. Chem. Phys.* **2009**, *131*, 14102.
- (55) Hunter, E. P.; Lias, S. G. *J. Phys. Chem. Ref. Data* **1998**, *27*, 413.
- (56) *CRC Handbook of Chemistry and Physics*, 80th ed.; Lide, D. R., Ed.; CRC Press, Inc.: New York, 1999; pp 10–175.
- (57) Karton, A.; Parthiban, S.; Martin, J. M. L. *J. Phys. Chem. A* **2009**, *113*, 4802.
- (58) Ng, C. Y.; Trevor, D. J.; Tiedemann, P. W.; Ceyer, S. T.; Kronebusch, P. L.; Mahan, B. H.; Lee, Y. T. *J. Chem. Phys.* **1977**, *67*, 4235.
- (59) Harder, E.; Anisimov, V. M.; Vorobyov, I. V.; Lopes, P. E. M.; Noskov, S. Y.; MacKerell, A. D., Jr.; Roux, B. *J. Chem. Theory Comput.* **2006**, *2*, 1587.
- (60) Harder, E.; Anisimov, V. M.; Whitfield, T.; MacKerell, A. D., Jr.; Roux, B. *J. Phys. Chem. B* **2008**, *112*, 3509.
- (61) Gao, J.; Habibollazadeh, D.; Shao, L. *J. Phys. Chem.* **1995**, *99*, 16460.
- (62) Gao, J. *J. Comput. Chem.* **1997**, *18*, 1062.
- (63) Xie, W.; Pu, J.; Mackerell, A. D., Jr.; Gao, J. *J. Chem. Theory Comput.* **2007**, *3*, 1878.
- (64) Xie, W.; Pu, J.; Gao, J. *J. Phys. Chem.* **2009**, *113*, 2109.
- (65) Voityuk, A. A.; Rösch, N. *J. Phys. Chem. A* **2000**, *104*, 4089.
- (66) Tang, K. T.; Toennies, J. P. *J. Chem. Phys.* **1984**, *80*, 3726.
- (67) Misquitta, A.; Stone, A. *Mol. Phys.* **2008**, *106*, 1631.
- (68) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *J. Chem. Phys.* **2010**, *132*, 154104.

Coupled Cluster in Condensed Phase. Part II: Liquid Hydrogen Fluoride from Quantum Cluster Equilibrium Theory

Christian Spickermann,^{†,§} Eva Perl,[†] Michael von Domaros,[†] Martin Roatsch,[†] Joachim Friedrich,[‡] and Barbara Kirchner^{*,†}

[†]Wilhelm-Ostwald-Institut für Physikalische und Theoretische Chemie, Universität Leipzig, Linnéstrasse 2, D-04103 Leipzig, Germany

[‡]Institute for Chemistry, Chemnitz University of Technology, Strasse der Nationen 62, 09111 Chemnitz, Germany

ABSTRACT: Treating the bulk phase with high-level ab initio methods, such as coupled cluster, is a nontrivial task because of the computational costs of these electronic structure methods. In this part of our hydrogen fluoride study we make use of the quantum cluster equilibrium method, which employs electronic structure input of small clusters and combines it with simple statistical mechanics in order to describe condensed phase phenomena. If no parameter adjustment is applied, then the lower quantum chemical methods, such as density functional theory in conjunction with the generalized gradient approximation, provide wrong results in accordance with the description of the strength of the interaction in the clusters. While density functional theory describes the liquid phase too dense due to overbinding of the clusters, the coupled cluster method and the perturbation theory at the complete basis set limit agree well with experimental observations. If we allow the two parameters in the quantum cluster equilibrium method to vary, then these are able to compensate the overbinding, thereby leading to very good agreement with experiment. Correlated methods in combination with small basis sets giving rise to too weakly bound clusters cannot reach this accuracy even if the parameters are flexible. Only at the complete basis set limit, the performance of the correlated methods is again excellent.

1. INTRODUCTION

Liquid hydrogen fluoride (HF) is a substance of importance for the chemical industry as well as academic research, e.g., its superacidic properties have been successfully employed for the investigation and clarification of reaction mechanisms.^{1–3} However, the structural as well as thermodynamic data available for liquid HF in literature has been rather scarce, which is not surprising considering the highly toxic and corrosive properties of this fluid. Indeed, it has been proposed to rather calculate its properties than to measure them.^{4,5} However, the accurate treatment of the liquid state in terms of computational chemistry is still a nontrivial task. The most obvious choice for the modeling of condensed systems lies in traditional molecular dynamics simulations, which account for the intermolecular interactions being important at high densities in terms of analytical potential functions fitted to reproduce macroscopic experimental data. Although long simulation times and large sample sizes can be realized in terms of this approach, the quantum mechanical character of the molecular system is most often completely neglected and effects, like cooperativity or spontaneous events, fall outside the reach of these kind of calculations. Ab initio molecular dynamics simulations can account for these deficiencies, but due to the large computational effort, simulation time as well as sample size are restricted considerably in these models, which makes the prediction of thermodynamical quantities very difficult for systems exhibiting many degrees of freedom. The quantum cluster equilibrium (QCE) theory⁶ circumvents the computational time bottleneck, i.e., these sampling problems, by modeling the liquid phase as a thermodynamic equilibrium of distinct cluster structures, for which an approximate but analytical partition function and thus a gateway to the thermodynamics of the condensed phase is available. Furthermore, the introduction of highly accurate post-Hartree–Fock electronic structure data, such as the coupled cluster method, is possible.

In cases of highly associated liquids, numerous successful applications of the QCE model in this field were demonstrated.^{7–21} However, liquid HF as one of the most generic associated systems has not been in the focus of the QCE approach so far, although there are several other computational studies dealing with this substance in literature.^{22–34} Most computational approaches toward HF in terms of static ab initio methods were mainly concerned with cooperative effects in isolated cluster structures,^{22–29} but several traditional as well as ab initio molecular dynamics studies on liquid HF have also been published, which mainly deal with structural aspects of the fluid phase.^{30–34}

Ab initio molecular dynamics simulations^{31,34} indicate a predominance of chain-like structures in liquid HF at ambient as well as supercritical states. In contrast, experimental studies do not definitely attest this observation but rather assume an equality between the structure of the solid (parallel zigzag chains) and the structure of the liquid.³⁵ This is also true in case of the elevated temperature regime, for which structural changes of the hydrogen-bonded species are indicated, but no prediction about the kind of these changes could be made.³⁵ Furthermore, important electron correlation effects have been neglected in these studies so far. Earlier experimental investigations also explicitly discuss the occurrence of cyclic species in the liquid phase and highlight that cooperative effects, which are assumed to play only a minor role in the liquid, could as well seriously affect the liquid phase structures of strongly associated liquids, such as water or HF.³⁶

The purpose of this article is to apply the coupled cluster method to the liquid phase of HF via the QCE approach. To the best of our

Received: January 31, 2011

Published: March 16, 2011

knowledge, the coupled cluster method has never been combined with the QCE approach. It has been used to derive ab initio force field parameters,³⁷ but it has not been used directly in condensed phase calculations.

The article is organized as follows. A short introduction to the QCE method is given in the next section, Section 2. After this, the investigated structures are presented and the appropriate binding energies are discussed. In Section 3, the results for the liquid phase are presented, and the influence of the electronic structure method is investigated. Finally, the behavior at the phase transition is presented.

2. THEORY

2.1. Quantum Cluster Equilibrium (QCE) Method. A full derivation of the QCE theory can be found elsewhere,^{8,9,13,38} with the most detailed description in refs 9 and 13. As laid out in these references, the basic idea of the QCE model is a thermodynamic equilibrium between different sized clusters and one corresponding reference monomer. The equilibrium equation thus reads

$$C_1 \rightleftharpoons \frac{C_2}{i(2)} \rightleftharpoons \dots \rightleftharpoons \frac{C_{\mathcal{P}}}{i(\mathcal{P})} \rightleftharpoons \dots \rightleftharpoons \frac{C_{\eta}}{i(\eta)} \quad (1)$$

Here $C_{\mathcal{P}}$ denotes a cluster of $i(\mathcal{P})$ monomer units up to $i(\eta)$ monomer units forming the largest cluster, and η represents the total number of clusters considered in the mixture. Using the relation between the molecular partition function $q_{\mathcal{P}}$ of the particle \mathcal{P} and the chemical potential $\mu_{\mathcal{P}}$ for which the same equilibrium holds, as in eq 1:

$$\mu_{\mathcal{P}} = -kT \ln \left(\frac{q_{\mathcal{P}}}{N_{\mathcal{P}}} \right) \quad (2)$$

followed by application of the particle conservation

$$N_A = N_1 + i(2)N_2 + \dots + i(\mathcal{P})N_{\mathcal{P}} + \dots + i(\eta)N_{\eta} \quad (3)$$

leads to the crucial step of the QCE calculation, i.e., an iterative cycle for the root finding of the population polynomial and the volume polynomial. Here k denotes the Boltzmann constant, T the temperature, N_A the Avogadro number, and $N_{\mathcal{P}}$ the particle number of particle \mathcal{P} . Using moles $n_{\mathcal{P}} = N_{\mathcal{P}}/N_A$ instead of particle numbers, the population polynomial for the monomer as reference species is given by the following expression:

$$0 = -1 + \sum_{\mathcal{P}=1}^{\eta} \left[\frac{i(\mathcal{P})q_{\mathcal{P}}N_A^{i(\mathcal{P})-1}}{q_1^{i(\mathcal{P})}} \right] n_1^{i(\mathcal{P})} \quad (4)$$

If there are cluster sizes which do not exist in the chosen cluster mixture, then the vanishing partition function $q_{\mathcal{P}}$ for this particular $i(\mathcal{P})$ value will ensure the correct form of the polynomial. The populations of all other clusters are completely determined in terms of the monomer population^{6,8} and allow the calculation of the canonical QCE partition function Q^{tot} according to

$$Q^{\text{tot}} = \prod_{\mathcal{P}=1}^{\eta} Q_{\mathcal{P}}(\{N_{\mathcal{P}}\}, V, T) = \prod_{\mathcal{P}=1}^{\eta} \frac{q_{\mathcal{P}}^{N_{\mathcal{P}}}}{N_{\mathcal{P}}!} \quad (5)$$

The molecular partition function $q_{\mathcal{P}}$ may be decomposed into its molecular degrees of freedom assuming that those are

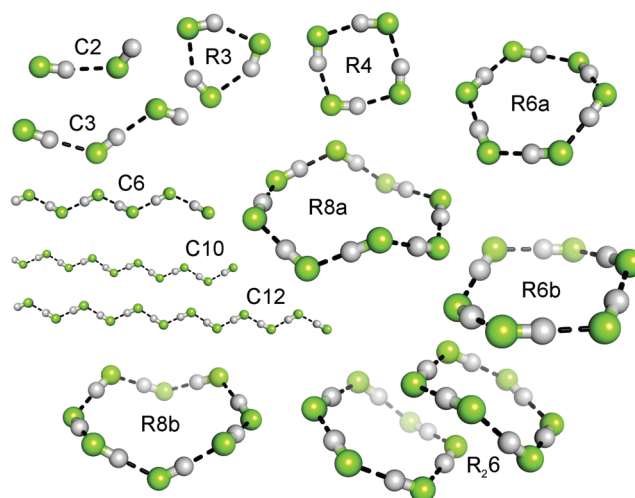


Figure 1. Ball-and-stick model of the investigated clusters as obtained by MP2/TZVP calculations. C denotes chain structures whereas R denotes cyclic structures. The R_{26} cluster is built up of two stacked six-membered rings.

independent of each other

$$q_{\mathcal{P}} = q_{\mathcal{P}}^{\text{tr}} \cdot q_{\mathcal{P}}^{\text{rot}} \cdot q_{\mathcal{P}}^{\text{vib}} \cdot q_{\mathcal{P}}^{\text{el}} \quad (6)$$

with the superscripts denoting translational, rotational, vibrational, and electronic contributions, respectively. By employing the relationship between pressure p and the partition function Q^{tot} , a third-order polynomial for the phase volume V can be derived, yielding

$$0 = -pV^3 + \left[kT \sum_{\mathcal{P}=1}^{\eta} N_{\mathcal{P}} + pV_{\text{excl}} \right] V^2 - \left[\sum_{\mathcal{P}=1}^{\eta} N_{\mathcal{P}} i(\mathcal{P}) a_{\text{mf}} \right] V + \left[\sum_{\mathcal{P}=1}^{\eta} N_{\mathcal{P}} i(\mathcal{P}) a_{\text{mf}} \right] V_{\text{excl}} \quad (7)$$

In this equation V_{excl} denotes an excluded volume term according to the population-weighted sum of the cluster volumes and a_{mf} an empirical cluster–cluster interaction parameter described in the following.

In order to treat interactions between clusters (intercluster interaction), a van der Waals-like meanfield term of the form

$$u_{\mathcal{P}}^{\text{int}} = u_{\mathcal{P}}^{\text{int}}(i(\mathcal{P}), V) = -a_{\text{mf}} \frac{i(\mathcal{P})}{V} \quad (8)$$

was suggested, with a_{mf} being a substance-specific scaling factor, and V the overall phase volume as obtained from the volume polynomial (eq 7).⁶ The energy term entering the electronic cluster partition function $q_{\mathcal{P}}^{\text{el}}$ is thus given as the sum of the intracuster and intercluster interactions, and $q_{\mathcal{P}}^{\text{el}}$ is obtained from

$$q_{\mathcal{P}}^{\text{el}} = \exp(-[E_{\mathcal{P}}^{\text{intra}} + u_{\mathcal{P}}^{\text{int}}]/kT) \quad (9)$$

where $E_{\mathcal{P}}^{\text{intra}}$ denotes the adiabatic intracuster interaction energy according to

$$E_{\mathcal{P}}^{\text{intra}} = E_{\mathcal{P}} - i(\mathcal{P})E_1 \quad (10)$$

Please note that besides a_{mf} the model depends on another parameter (b_{sv}). It scales the cluster volume estimates according to the sum of van der Waals spheres so that it can be excluded for

Table 1. Adiabatic Interaction Energies $E_{\mathcal{L}}^{\text{intra}}$ According to eq 10 for All Investigated Clusters and Different Electronic Structure Methods^a

cluster	$i(\mathcal{L})$	CCSD(T)		MP2			B-P86	PBE
		CBS	CBS	QZVP*	QZVP	TZVP	TZVP	TZVP
C2	2	-19.48	-19.19	-18.73	-18.73	-18.54	-19.23	-22.92
C3	3	-45.13	-44.60	-43.58	-43.57	-42.98	-45.74	-53.13
R3	3	-65.12	-63.99	-62.06	-62.06	-56.80	-69.38	-78.39
R4	4	-119.01	-118.43	-115.12	-115.12	-106.36	-131.97	-145.17
C6	6	–	–	–	–	-129.88	-145.05	-163.15
R6a	6	-198.72	-198.85	-193.09	-193.04	-183.78	-223.57	-244.80
R6b	6	-199.05	-199.19	-193.41	-193.41	-183.95	-224.34	-245.55
R8a	8	-268.81	-269.00	-261.30	-260.82	-250.46	-303.98	-331.77
R8b	8	-269.11	-269.40	-261.92	-261.81	-250.84	-305.27	-333.54
C10	10	–	–	–	–	-254.41	–	–
C12	12	–	–	–	–	-317.70	-368.30	–
R ₂ 6	12	-415.07	-412.41	-397.82	-397.82	-375.71	-447.04	-499.42

^a All energies in kJ/mol. Some results are missing because the corresponding geometries are no minimum structures.

translation.⁸ Both parameters a_{mf} and b_{xv} are determined by a parameter optimization technique (PO) which works as follows: Different combinations of values for a_{mf} and b_{xv} enter a QCE calculation yielding different isobars $V_c(T)$. For each of these isobars an error norm $\|\Delta V\|$ with respect to an experimental reference isobar $V_r(T)$ is determined according to

$$\|\Delta V\| = \sqrt{\sum_T (V_c(T) - V_r(T))^2} \quad (11)$$

with the summation running over all temperatures. The combination resulting in the lowest error norm is taken to be most reliable.¹⁸

In order to investigate the impact of certain clusters or cluster motifs on the liquid phase and to identify destabilizing structures, a cluster set optimization procedure (CSO) is additionally carried out, i.e., clusters giving populations below a certain threshold value over the whole temperature range are systematically left out and taken into account, and thereby for each permutation, a new cluster set is obtained. For each of these sets a parameter optimization is carried out. Finally, the cluster set together with its optimized parameters that yields the lowest error norm is taken to be the best result.

The QCE calculations were performed employing our own software, i.e., the PEACEMAKER code.³⁸ This code is freely available from <http://www.uni-leipzig.de/~peacemaker/>.

3. RESULTS

In order to perform QCE calculations we need a set of clusters with characteristic topologies similar to a basis set in quantum chemical calculations. The clusters building up the set should represent certain motifs of the investigated phase rather than portray the phase in analogy to a complete basis set in quantum chemistry. In case of HF we chose different sized ring and chain structures. Furthermore we included the R₂6 cluster which consists of two stacked ring structures. Figure 1 shows all clusters from which certain examples were used as input in the following QCE calculations.

3.1. Interaction Energies. The interaction energies which enter eq 10 for the different clusters are calculated with different

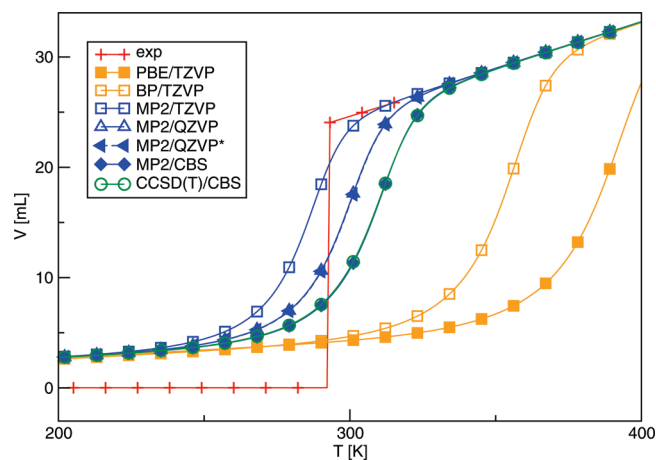


Figure 2. Isobars from different quantum chemical methods without adjusting the parameters, i.e., the QCE⁽⁰⁾ model.

electronic structure methods, see table 1. Computational details can be found in ref 39. Concerning the basis set superposition error, ref 40 shows that the CP-corrected energies converge from above to the complete basis set limit, whereas the noncorrected values converge from below to the complete basis set limit with increasing basis sets. Therefore, the energies applied to the QCE calculations are the averaged values of CP-corrected and uncorrected energy.

As one can see from Table 1, energies as obtained by the density functional theory (DFT) approaches are smaller, and thus clusters are suggested to be more stable than those obtained by the MP2 and the CCSD(T) approach. Regarding the MP2 calculations, one observes a strong basis set dependence with slightly increasing stability of the clusters when enlarging the basis set applied. A similar trend is observed when improving the level of theory from MP2 to CCSD(T) except for the six- and eight-membered rings, where results are comparable for both methods. For further discussion of energies see ref 39.

3.2. Condensed Phase Calculations: Liquid. In the following we will use the quantum chemically obtained cluster energies and

Table 2. Optimized QCE Parameters (a_{mf}/b_{xv}) and Error Norm (in mL) of the Different Methods Employed in This Study^a

method	PO			CSO			
	a_{mf}	b_{xv}	$\ \Delta V\ $	a_{mf}	b_{xv}	$\ \Delta V\ $	cluster
PBE/TZVP	0.0253	1.361	0.226	0.0253	1.361	0.226	1, R4, R6a, R6b, R8a
B-P86/TZVP	0.0239	1.372	0.246	0.0238	1.370	0.240	1, C6, R6a, R6b, R8a
MP2/TZVP	0.0672	1.663	3.888	0.0671	1.663	3.886	1, R3, C3, R4, C6, R6a, R6b, R8a
MP2/QZVP	0.0349	1.495	2.280	0.0346	1.492	2.240	1, R3, C3, R4, R6a, R6b, R8a
MP2/QZVP*	0.0348	1.494	2.265	0.0344	1.488	2.224	1, C3, R4, R6a, R6b, R8a
MP2/CBS	0.0255	1.357	0.261	0.0253	1.354	0.254	1, R3, R6a, R6b, R8a
CCSD(T)/CBS	0.0256	1.359	0.265	0.0253	1.354	0.250	1, R6a, R6b, R8a

^a First block: parameter optimization (PO), and second block: cluster set optimization (CSO). Stepsize in PO $a_{mf} = 0.0001$ and $b_{xv} = 0.001$. Population threshold for CSO 2%.

harmonic frequencies from ref 39 in order to calculate thermodynamic data of the condensed phase. In order to compare the different levels of electronic structure theory, the smaller clusters 1, C2, C3, C6, R3, R4, R6a, R6b, and R8a are considered only, and the R8b, C10, C12, and R₂6 clusters are neglected within the first part and will be taken into account in the latter part of this article. Furthermore, to gain insight into the performance of the electronic structure methods, we first apply the quantum cluster equilibrium theory without adjusting the parameters. Thus, the next section shows the obtained isobars from the QCE⁽⁰⁾ model.

3.2.1. Evaluation of Methodology: QCE⁽⁰⁾. In this section we will only show the isobars calculated from applying $a_{mf} = 0$ and $b_{xv} = 1$, i.e., the QCE⁽⁰⁾ model. Thus, there is no interaction between clusters, and the cluster volume is taken into account as the unscaled van der Waals estimate. From the behavior of the obtained isobars we can estimate the influence of the electronic structure method applied, see Figure 2. The orange curves show the DFT results, the blue ones mark MP2 values, and in green we give the CCSD(T) data. All curves show a phase transition which is quite fascinating, keeping in mind that all input stems from static quantum chemical calculations of isolated particles. By setting the QCE parameters to 0 and 1, respectively, a cluster gas is represented. The phase transition as obtained by the DFT methods occurs at higher temperatures than measured by experiment. The calculated volumes are too small which can be explained in terms of “overbinding” as indicated in Table 1. Concerning MP2 values in combination with Ahlrichs basis sets, we observe an opposite trend: At the triple- ζ basis set the temperature of the phase transition is observed to be smaller than measured by experiment. In this case too large volumes indicate “underbinding”. This is reduced with increasing basis sets. The phase transition temperature at the complete basis set limit (MP2 and CCSD(T)) is found to be reliable. It is clear that this observation has an impact on the condensed phase. The question is to what extent the QCE parameters can compensate these shortcomings of some of the electronic structure methods when they are optimized.

There is one other aspect worth mentioning. MP2/QZVP*, MP2/CBS, and CCSD(T)/CBS all rely on the same harmonic frequencies and geometries. Therefore, the influence of the accurate calculation of the energy is clearly recognizable. Both curves obtained at the CBS are based on very sophisticated energetics, see ref 39, which thereby can be identified as a necessary prerequisite for an accurate prediction of the phase transition point in the frame of the QCE⁽⁰⁾ approach.

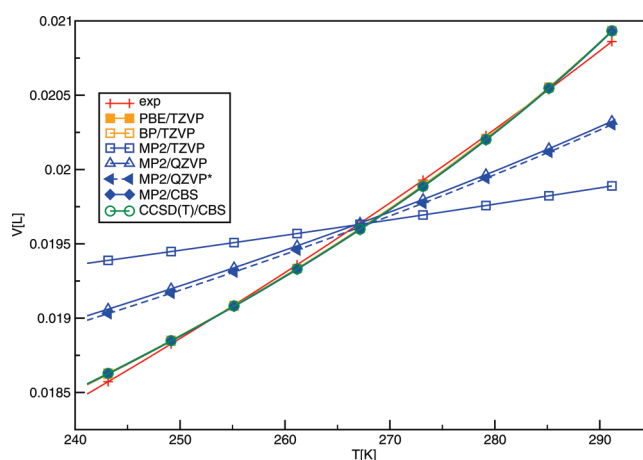


Figure 3. Liquid phase isobars from different quantum chemical methods by adjusting the two parameters a_{mf} and b_{xv} .

3.2.2. Liquid Phase: Isobars and Thermal Expansion Coefficient. We now turn to the optimized QCE model in the liquid phase, i.e., to those calculations in which we adjusted the a_{mf} and b_{xv} values in order to describe the real liquid. Table 2 shows the optimized parameters (PO) a_{mf} and b_{xv} together with the appropriate error norm $\|\Delta V\|$ as well as the result of the cluster set optimization (CSO) with regard to the experimental volume for each method applied.

Table 2 shows that overbinding as observed in case of the DFT calculations can be compensated by optimizing QCE parameters, as here the obtained error norms are comparable to those of the post-Hartree–Fock methods at the complete basis set limit. The MP2 values combined with Ahlrichs basis sets in contrast show deviations which are larger by factor 10. In order to compensate the smaller interaction energies the MP2/TZVP combination provides a larger meanfield parameter a_{mf} and free volume correction b_{xv} . Thus, one can assume that underbinding in intracuster energies demand to be balanced by larger intercluster interactions in order to reproduce liquid phase behavior. However, the compensation of the underbinding in terms of the meanfield contribution is not as accurate as the overbinding correction for DFT, as can be seen in the large $\|\Delta V\|$ values obtained for the finite basis set MP2 methods. As a verification, we also adjust a QCE calculation to PBE/TZVP energies and MP2/QZVP* frequencies and geometries. The obtained values are $a_{mf} = 0.0246$, $b_{xv} = 1.3680$, and $\|\Delta V\| = 0.236$. Thus overbinding can easily be corrected which is reflected in the smaller

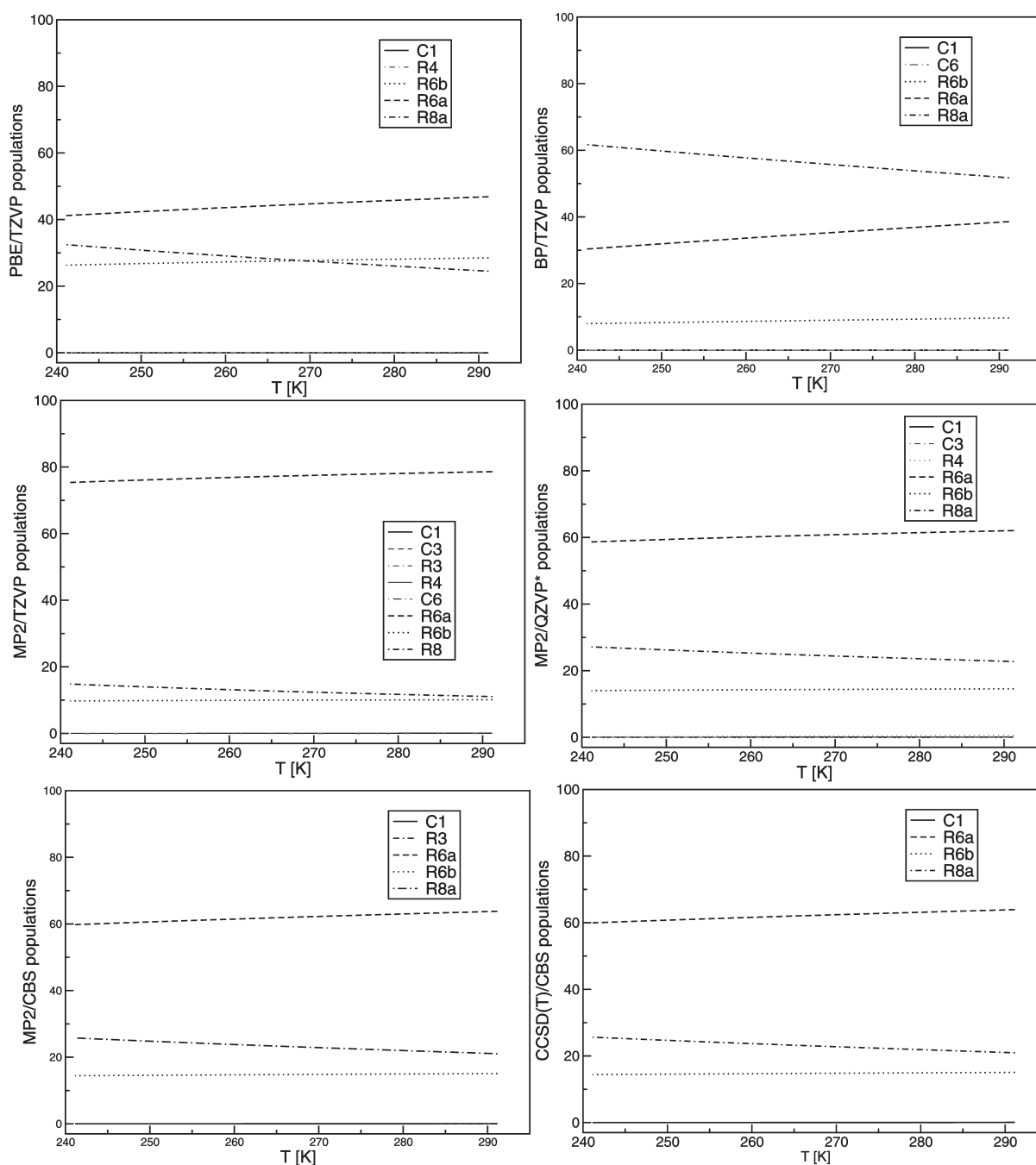


Figure 4. Populations for the different electronic structure methods as obtained from the CSO QCE calculations.

a_{mf} value compared to the pure MP2/QZVP* or MP2/CBS or CCSD(T)/CBS calculations.

Previously, we found that optimizing the cluster set, i.e., excluding some clusters can lead to tremendously improved results thus addressing a destabilizing nature to some clusters.¹⁹ This does not apply to HF, where we found only minor differences leading to improved $\|\Delta V\|$ as small as 0.01 mE_h. However, one important result of the CSO procedure is that the six- and eight-membered ring structures are selected independent of the level of theory and the basis set applied, which will be discussed later.

The isobars from the above-discussed PO calculations are depicted in Figure 3. As observed already from Table 2, PBE/TZVP, B–P86/TZVP, MP2/CBS, and CCSD(T)/CBS show

excellent agreement with experiment, while the other MP2 values are much more departing from experiment. The large basis set dependency for MP2 is also apparent from Figure 3; compare blue curves with different symbols. These results also show that in a QCE calculation the parameter optimization is unable to account for all deficiencies of the electronic structure methods. While overbinding can be compensated, even slight underbinding is not correctable by a sophisticated choice of parameters.

In order to learn about the topology of liquid phase structures we also show the populations as obtained by the cluster set optimization in Figure 4.

Independent of the methodology, only the ring clusters containing six and eight monomer units contribute significantly. For all

methods the R6a population increases with increasing temperature, and we observe the highest individual cluster population for the R6a cluster for all methods except for the B–P86/TZVP approach. The cluster R8a also contributes significantly, and its population decreases for all methods with increasing temperature. It is clear from simple rules of thumb that the more hydrogen bonds are formed, the more stable ring clusters are in the gas phase. However, the larger the ring clusters are the more they resemble chain clusters which was stated by Hammes-Schiffer and co-workers.⁴¹ Therefore, it is reasonable that the larger clusters show higher populations than the smaller clusters which will be further investigated in Section 3.2.3.

Figure 5 shows the temperature dependency of the thermal expansion coefficient. Please note that this quantity is calculated according to

$$\alpha = \frac{1}{V} \left(\frac{\partial V}{\partial T} \right) \quad (12)$$

and, thus, does not directly depend on the partition function but on the volume and its first derivative. This is the reason why those methods able to accurately reproduce the phase volume yield values for the thermal expansion coefficient are in good agreement with experimental data.

3.2.3. Extending the Motifs: Larger Chain Clusters. Turning now to QCE calculations with larger clusters, we show the results of the B–P86/TZVP data including the R8b, C12, and R₂6 and of the MP2/TZVP input data including the R8b, C10, C12, and R₂6 clusters, see Table 3. The choice of these two methods is related to the fact that for other electronic structure methods, the C10 and C12 are no minimum structures. The a_{mf} value is decreased for the B–P86/TZVP input data and increased in case

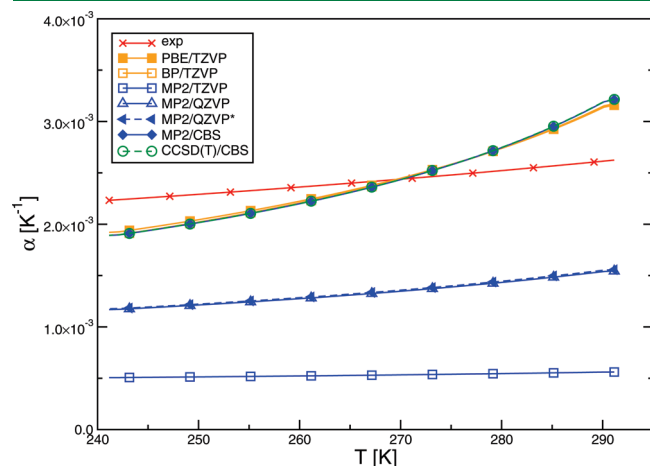


Figure 5. Thermal expansion coefficient α for different quantum chemical methods as obtained from the CSO QCE calculations. Experimental data from ref 42.

Table 3. Optimized QCE Parameters (a_{mf}/b_{xv}) and Error Norm (in mL) of the Different Methods Employed in This Study^a

method	PO			CSO			
	a_{mf}	b_{xv}	$\ \Delta V\ $	a_{mf}	b_{xv}	$\ \Delta V\ $	cluster
B-P86/TZVP	0.0225	1.366	0.255	0.0225	1.366	0.255	C1, R4, R6a, R6b, C6, R8a, R8b
MP2/TZVP	0.0674	1.668	3.904	0.0672	1.667	3.889	C1, C3, R3, R4, C6, R6a, R6b, R8a, R8b, R ₂ 6

^a First block: parameter optimization (PO), and second block: cluster set optimization (CSO). Stepsize in PO $a_{mf} = 0.0001$ and $b_{xv} = 0.001$. Population threshold for CSO 2%.

of the MP2/TZVP. However, the accuracy is not improved but slightly worse than for the smaller cluster set.

This is also the case if instead of a PO calculation a CSO calculation is chosen, see Table 3. The results are slightly worse than for the CSO with the smaller cluster set. Regarding the population, almost nothing changed compared to calculations applying the smaller cluster set, as the dominant species still consist of six-membered and eight-membered rings (populations not shown here).

3.3. Phase Transition: Isobars. In this section we consider the results up to the boiling point with optimized parameter values from a new PO calculation. In this case the reference isobar only consists of two values at the temperatures 292.15 and 293.15 K, the highest temperature in the liquid phase and the lowest

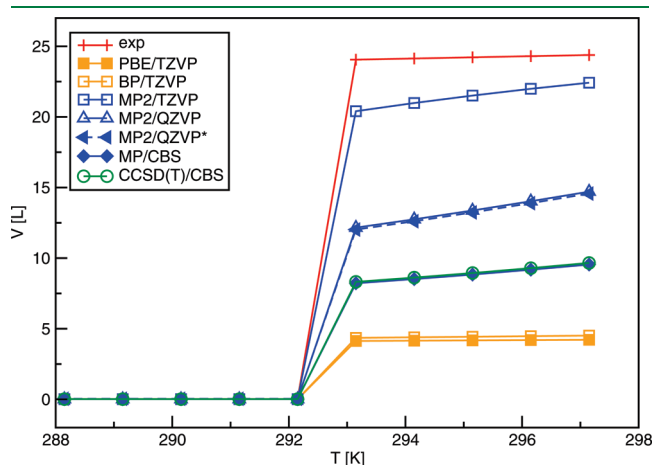


Figure 6. Liquid–gas phase transition isobars from different quantum chemical methods by optimizing the two parameters a_{mf} and b_{xv} for the phase transition.

Table 4. Optimized QCE Parameters (a_{mf}/b_{xv}) and Error Norm (in L) of the Different Methods Employed in This Study^a

method	PO		
	a_{mf}	b_{xv}	$\ \Delta V\ $
PBE/TZVP	0.0088	1.009	19.920
B–P86/TZVP	0.0092	1.005	19.698
MP2/TZVP	0.0419	1.001	3.650
MP2/QZVP	0.0283	1.153	11.900
MP2/QZVP*	0.0754	3.103	12.033
MP2/CBS	0.0169	1.005	15.834
CCSD(T)/CBS	0.0171	1.005	15.733

^a Stepsize in PO $a_{mf} = 0.0001$ and $b_{xv} = 0.001$.

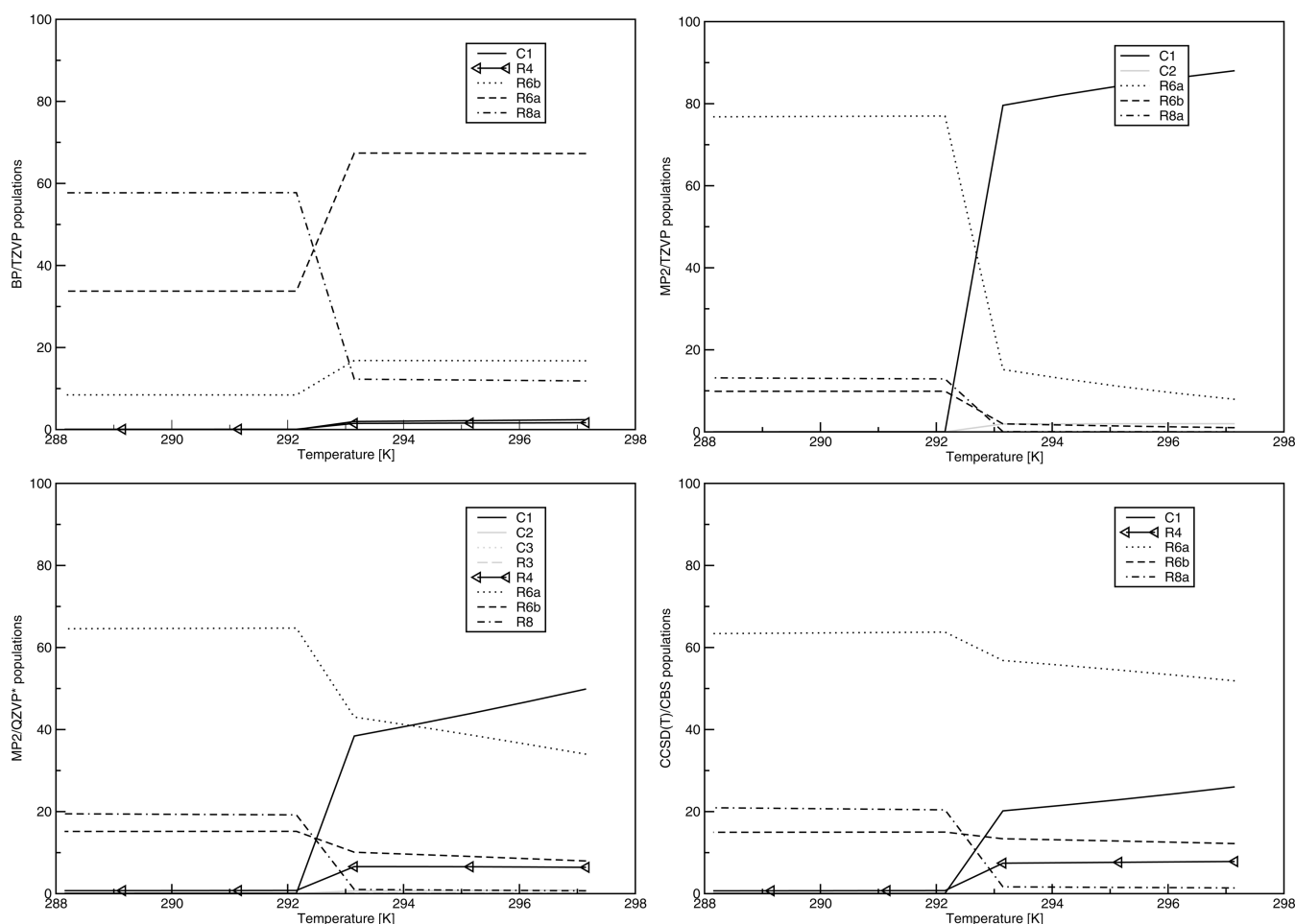


Figure 7. Liquid–gas phase transition populations for different quantum chemical methods by optimizing the two parameters a_{mf} and b_{xv} for the phase transition.

temperature in the gaseous phase, respectively. Of course, the selection of adjustment temperatures might be enhanced, however, for a first indication of the performance at the boiling point, this two-point fit is sufficient enough. The corresponding results are shown in Figure 6, and the obtained parameters from this two-point adjustment are given in Table 4. Please note that in Table 4 the squared deviations are given in L and not in mL.

The difficulty in reproducing the phase transition is apparent by the fact that (in the gas phase) now different error cancellations play a role, and therefore different quantum chemical methods perform best contrary to the behavior in the bulk liquid phase. From Figure 6 and Table 4 we recognize that MP2/TZVP with its smallest interaction energies yields the highest accuracy. While DFT performs worst, the numbers from the best electronic structure method lie in between. Interestingly the excluded volume is almost not scaled ($b_{xv} = 1$) in order to account for the gas phase. Please note that we did not allow this parameter to shrink beyond 1.0 in this particular optimization.

In Figure 7 we show the cluster populations as they occur in our QCE calculations for four selected electronic structure methods. Please note that the neglected electronic structure cases give similar results to those methods which they match regarding the isobars. Obviously, the good performance of MP2/TZVP is due to the population of the monomer species, and the bad performance of DFT is due to the missing occupancy of the monomer species. Interestingly, in the region of the boiling point

the vapor phase consists still of a variety of different clusters, for example both six-membered rings and even the eight-membered rings contribute to some extent. McGrath et al. have recently investigated the vapor of HF from ab initio Monte Carlo simulations.⁴³ Although the authors worked at much higher temperature, they found the prominence of monomeric species as well, next to some populations of a six-membered ring.⁴³

4. CONCLUSION

We presented the first quantum cluster equilibrium (QCE) calculations for the liquid phase based on highly accurate CCSD(T) computations of the underlying cluster structures. Furthermore the coupled cluster method and the second order Møller–Plesset perturbation theory (MP2) were applied at the complete basis set limit. In order to compare the outcome of the QCE calculations, we also applied other electronic structure data as input. These were obtained from density functional theory (DFT), namely B-P86/TZVP and PBE/TZVP, and from the MP2 method in combination with different Ahlrichs basis sets. The interaction energies are such that DFT provides strongest hydrogen bonds, MP2/TZVP weakest interaction energies, and the CCSD(T) and MP2 at the complete basis set limit values in between in high agreement with experimental data. The QCE⁽⁰⁾ calculations were carried out without taking additional interaction or volume scaling into account. For all

electronic structure methods this led to isobars with boiling points more or less away from experimental values. While the DFT methods provided too small volumes and therefore too large densities showing the well-known behavior of overbinding, the MP2/TZVP values were slightly too large, and the data obtained from the complete basis set limit calculations were very close to the experimental boiling point.

In order to model a realistic liquid, the two parameters were adjusted, and it was found that DFT and the correlated electronic structure methods at the complete basis set limit agreed best with the experimental isobar. For the MP2 combinations with the smaller basis sets no such agreement with experiment could be achieved despite the fact that two parameters are used. From the populations we learned that an important topology lies in the six- and eight-ring clusters. While the former is increasing with higher temperature, the latter is decreasing. Inclusion of larger chain clusters did not change the picture significantly, still the six- and eight-membered rings played the dominant role in order to reproduce the liquid phase.

We also provided a first estimate of the performance at the phase transition. Here different error cancellations played a role. For example, the weakest bound structures (MP2/TZVP) were now leading to the best agreement with the experimental curve, because the gas phase can be modeled more accurately by those methods which yield less stable hydrogen bonds. In all cases we found a variety of clusters with small populations at the vapor phase in agreement with literature.⁴³ In further studies we will examine these issues in more detail and investigate other cluster structures.

To summarize, we can say that in order to arrive at a consistent multiscale condensed phase description based on the resolution of the electronic structure, the accurate treatment of the underlying electronic structure problem is necessary. However, if one is only interested in good agreement with experiment, a pragmatic point of view in which error cancellation is taken into account is possible if overbound cluster energies are used for the liquid phase and underbound energies for the gas phase.

AUTHOR INFORMATION

Corresponding Author

*E-mail: bkirchner@uni-leipzig.de.

Present Addresses

⁵Lehrstuhl für Anorganische Chemie 2, Organometallics and Materials Chemistry, Ruhr-Universität Bochum, Universitätsstrasse 150, D-44780 Bochum.

ACKNOWLEDGMENT

This work was supported by the DFG, in particular by the projects KI-768/4-1 and KI-768/4-2 from the ERA-chemistry, KI-768/5-1, and KI-768/5-2 SPP-IL program and the KI-768/7-1 project. Computer time from the RZ Leipzig is gratefully acknowledged.

REFERENCES

- (1) Fielding, H.; Lee, B. *Chem. Br.* **1978**, *14*, 173.
- (2) O'Donnell, T. A. *Superacids and Acidic Melts as Inorganic Chemical Reaction Media*; VCH Publishers: New York, 1992; pp 41–51.
- (3) Olah, G. A. *Angew. Chem., Int. Ed. Engl.* **1995**, *34*, 1393–1405.
- (4) McLain, S. E.; Benmore, C. J.; Siewenie, J. E.; Urquidí, J.; Turner, J. F. C. *Angew. Chem., Int. Ed. Engl.* **2004**, *43*, 1952–1955.
- (5) Visco, D. P.; Kofke, D. A. *Fluid Phase Equilib.* **1999**, *158*, 37–47.
- (6) Weinhold, F. *J. Chem. Phys.* **1998**, *109*, 367–372.
- (7) Weinhold, F. *J. Chem. Phys.* **1998**, *109*, 373–384.
- (8) Kirchner, B. *J. Chem. Phys.* **2005**, *123*, 204116.

- (9) Kirchner, B. *Phys. Rep.* **2007**, *440*, 1–111.
- (10) Ludwig, R.; Weinhold, F. *J. Chem. Phys.* **1999**, *110*, 508–515.
- (11) Ludwig, R.; Weinhold, F.; Farrar, T. C. *J. Chem. Phys.* **1995**, *103*, 3636–3642.
- (12) Matisz, G.; Fabian, W. M. F.; Kelterer, A.-M.; Kunsagi-Mate, S. *J. Mol. Struct. (THEOCHEM)* **2010**, *956*, 103–109.
- (13) Kirchner, B.; Spickermann, C.; Lehmann, S. B. C.; Perlt, E.; Langner, J.; von Domaros, M.; Reuther, P.; Uhlig, F.; Kohagen, M.; Brüssel, M. *J. Comp. Phys. Comm.* **2011** submitted.
- (14) Borowski, P.; Jaroniec, J.; Janowski, T.; Woliński, K. *Mol. Phys.* **2003**, *101*, 1413–1421.
- (15) Wendt, M. A.; Weinhold, F.; Farrar, T. C. *J. Chem. Phys.* **1998**, *109*, 5945–5947.
- (16) Ludwig, R.; Behler, J.; Klink, B.; Weinhold, F. *Angew. Chem., Int. Ed. Engl.* **2002**, *41*, 3199–3202.
- (17) Song, H.-J.; Xiao, H.-M.; Dong, H.-S.; Huang, Y.-G. *J. Mol. Struct. (THEOCHEM)* **2006**, *767*, 67–73.
- (18) Spickermann, C.; Lehmann, S. B. C.; Kirchner, B. *J. Chem. Phys.* **2008**, *128*, 244506.
- (19) Lehmann, S. B. C.; Spickermann, C.; Kirchner, B. *J. Chem. Theory Comput.* **2009**, *5*, 1640–1649.
- (20) Lehmann, S. B. C.; Spickermann, C.; Kirchner, B. *J. Chem. Theory Comput.* **2009**, *5*, 1650–1656.
- (21) Lenz, A.; Ojamae, L. *J. Chem. Phys.* **2009**, *131*, 134302.
- (22) Maerker, C.; v. R. Schleyer, P.; Liedl, K. R.; Ha, T. K.; Quack, M.; Suhm, M. A. *J. Comput. Chem.* **1997**, *18*, 1695–1719.
- (23) Quack, M.; Stohner, J.; Suhm, M. A. *J. Mol. Struct.* **2001**, *599*, 381–425.
- (24) Buth, C.; Paulus, B. *Phys. Rev. B* **2006**, *74*, 045122.
- (25) Wierzchowski, S. J.; Fang, Z. H.; Kofke, D. A.; Tilson, J. L. *Mol. Phys.* **2006**, *104*, 503–513.
- (26) Li, J. J. *Theor. Comp. Chem.* **2006**, *5*, 187–196.
- (27) Rinçon, L.; Almeida, R.; Garca-Aldea, D.; Diez y Riega, H. *J. Chem. Phys.* **2001**, *114*, 5552–5561.
- (28) Guedes, R. C.; do Couto, P. C.; Cabral, B. J. C. *J. Chem. Phys.* **2003**, *118*, 1272–1281.
- (29) Sangster, M. J. *Phys. Chem. Solids* **1974**, *35*, 195–200.
- (30) Klein, M. L.; McDonald, I. R. *J. Chem. Phys.* **1979**, *71*, 298–308.
- (31) Röthlisberger, U.; Parrinello, M. *J. Chem. Phys.* **1997**, *106*, 4658–4664.
- (32) Jedlovsky, P.; Vallauri, R. *J. Chem. Phys.* **1997**, *107*, 10166–10176.
- (33) Valle, R. G. D.; Gazzillo, D. *Phys. Rev. B* **1999**, *59*, 13699–13706.
- (34) Kreitmair, M.; Bertagnolli, H.; Mortensen, J. J.; Parrinello, M. *J. Chem. Phys.* **2003**, *118*, 3639–3645.
- (35) Pfeleiderer, T.; Waldner, I.; Bertagnolli, H.; Tölheide, K.; Fischer, H. E. *J. Chem. Phys.* **2000**, *113*, 3690–3696.
- (36) Deraman, M.; Dore, J.; Powles, J.; Holloway, J. H.; Chieux, P. *Mol. Phys.* **1985**, *55*, 1351–1367.
- (37) Huber, H.; Dyson, A. J.; Kirchner, B. *Chem. Soc. Rev.* **1999**, *28*, 121–133.
- (38) Kirchner, B.; Spickermann, C. *PEACEMAKER*, v1.4; University of Bonn, Institute of Physical and Theoretical Chemistry, University of Leipzig, Wilhelm-Ostwald Institute of Physical and Theoretical Chemistry Bonn: Leipzig, Germany, 2008; <http://www.uni-leipzig.de/~quant/index.html/> and <http://www.uni-leipzig.de/~peacemaker/>. Both accessed March 1, 2011.
- (39) Friedrich, J.; Perlt, E.; Roatsch, M.; Spickermann, C.; Kirchner, B. *J. Chem. Theory Comput.* **2011**; doi: 10.1021/ct100131c.
- (40) Halkier, A.; Klopper, W.; Helgaker, T.; Jørgensen, P.; Taylor, P. R. *J. Chem. Phys.* **1999**, *111*, 9157.
- (41) Swalina, C.; Wang, Q.; Chakraborty, A.; Hammes-Schiffer, S. *J. Phys. Chem. A* **2007**, *111*, 2206–2212.
- (42) Yaws, C. L. *Chemical Properties Handbook*; McGraw-Hill: New York, 1999; p 639.
- (43) McGrath, M. J.; Ghogomu, J. N.; Mundy, C. J.; Kuo, I.-F. W.; Siepmann, J. I. *Phys. Chem. Chem. Phys.* **2010**, *12*, 7678–7687.

Effect of Triples to Dipole Moments in Fock-Space Multireference Coupled Cluster Method

Lalitha Ravichandran, Nayana Vaval, and Sourav Pal*

Physical Chemistry Division, National Chemical Laboratory, Pune-411008, India

ABSTRACT: In this paper, we present the new implementation of partial triples for the dipole moment of doublet radicals in Lagrangian formulation of Fock-space multireference coupled cluster (Λ -FSMRCC) response method. We have implemented a specific scheme of noniterative triples, in addition to singles and doubles schemes, which accounts for the effects appearing at least at the third order in dipole moments. The method is applied to the ground states of OH, OOH, HCOO, CN, CH, and PO radicals.

I. INTRODUCTION

Single reference coupled cluster (SRCC)^{1–7} has been accepted as the state-of-the-art method for the electronic structure calculations. It has been successfully implemented for the energy, gradients, molecular properties, and potential energy surfaces.^{8–18} In general, SRCC introduces dynamic electron correlation, which keeps electrons apart. It is well-known that triple excitations in SRCC contribute to the energy from fourth order onward. So far a different version of the SRCC method with full or partial inclusion of triples with increasing precision has been developed^{19–24} for energy. The noniterative triples are routinely used for high accuracy with an economical treatment of triples. The full inclusion of triples is expensive, though in the SRCC it has been implemented by Bartlett and co-workers²¹ for energy. The perturbative treatment of the quadruple excitations has also been attempted^{23,25} in single reference context. However there are cases which involve several configurations which make nearly equal contribution (quasidegenerate) to the exact wave function, i.e., bond-breaking situations in the ground or excited states, where SRCC fails. The restricted open (RO) shell-based CC method,²⁶ which uses a linear operator have been successful in describing the quasidegenerate cases. Though in single reference framework, selected triple and quadruple level excitations^{27,28} have been considered for quasidegenerate cases,^{29,30} multideterminantal or multireference coupled cluster (MRCC) methods have emerged as the methods of choice to take into account the quasidegenerate molecular systems.³¹ Among the multireference methods, the effective Hamiltonian-based^{32–34} MRCC methods provide multiple roots via diagonalization of the effective Hamiltonian within the model space. This subclass mainly spans two approaches: namely the Hilbert-space (HS) MRCC^{35–37} and Fock-space (FS) MRCC.^{39–44} HSMRCC assumes different vacua for different configurations in the model space with same number of electrons and a state universal wave operator to introduce the model space to virtual space excitations. The method is suitable for potential energy surface (PES) studies³¹ and situations involving curve crossing. However for PES, the state-selective MRCC method developed by Mukherjee and co-workers^{45,46} has been found to be more attractive in recent years from the point of view of circumventing the problem of intruder states. The

FSMRCC theory was originally formulated by Kutzelnigg,³⁸ Mukherjee,^{39–41} and Lindgren,⁴² and applications to atoms were made by Kaldor and co-workers.⁴³ The FSMRCC applications to molecules were performed by Pal et al.⁴⁴ FSMRCC is based on the concept of a common vacuum and assumes a valence universal wave operator to describe the various states, which are generated by addition and/or removal of electrons to/from the common vacuum, usually the closed-shell RHF configuration. FS methods are suitable for the difference energy calculations and thus describe ionized, electron attached, or excited states of a closed-shell system. However, both these methods suffer from the problem of intruder states. This problem can be avoided with the help of intermediate Hamiltonian-based³³ formulation both in Fock^{47,48} as well as Hilbert spaces.^{49,50}

The important feature of both the multireference formulations is their size extensivity. On the other hand, equation of motion (EOMCC)^{51–57} or linear response CC (CCLR) methods^{58–60} use a linear operator for an excited state but an exponential operator for the ground state. The EOMCC method has been extensively developed for ionized,^{53,54} electron-attached,⁵⁵ and excited-state⁵¹ problems. The similarity transformed EOMCC method (STEOMCC), which is size extensive, was developed by Nooijen and co-workers.^{61,62} For one valence problem, EOMCC and FSMRCC are equivalent. However, such equivalence breaks down for excited state. EOMCC contains certain unlinked diagrams which are associated with charge-transfer separability.⁶¹ The spin-flip EOMCC method has also been introduced as a clever way to describe the multireference states.⁶³ The symmetry-adapted cluster expansion configuration interaction (SACCI)^{64,65} and method of moments coupled cluster (MMCC)^{66,67} have also been successful in describing some quasidegenerate problems. There are several implementations of the full and partial inclusion of the triples within the Fock-space^{68–71} MRCC. Pal and co-workers included noniterative triples for ionization potential^{68,69} and excitation energies,⁷⁰ within a FSMRCC scheme, and Bartlett and co-workers included full triples correction for excitation energies.⁷¹ The full triples correction to excitation energies in intermediate Hamiltonian FSCC has been pursued

Received: July 14, 2010

Published: March 10, 2011

currently by Musial et al.⁷² The inclusion of iterative and non-iterative triples in EOMCC^{73–77} and state-selective approaches^{78,79} for energy calculations has also been attempted. The perturbative triples corrections to EOM-IP-CCSD was introduced by Stanton and Gauss.⁸⁰ Recently, perturbative triples correction to EOM-EA-CCSD has been done by Manohar et al.⁸¹ The selected set of triples defined through the active orbitals in EOMCCSD (EOMCCSDt) has also been attempted.⁸² Recently, Krylov et al. employed the noniterative perturbative triples correction to the spin-flip EOMCC (SF-EOMCC) method for excitation energies.⁸³ The noniterative energy corrections to MMCC for excitation energy has been achieved by Piecuch et al.⁸⁴

Formulation of energy derivatives using multiroot CC methods is a challenging task. The response theory has been a valuable theoretical tool to study molecular properties.^{85,86} Along the lines of nonvariational CC (NVCC) response approach of Monkhorst, a response approach was developed for FSMRCC formulation^{87,88} and implemented for FSMRCC-based dipole moments of various ionized/electron-attached states as well as excited states. This method explicitly calculates the first derivatives of all cluster amplitudes^{89,90} and thus was not a satisfactory approach. Extending the idea of the Lagrange multipliers for the specific root of the effective Hamiltonian, Pal and co-workers developed the response approach within the MRCC framework (Λ -MRCC). This approach was formulated for the Hilbert-space⁹¹ as well as Fock-space⁹² MRCC methods. This formulation is very general and can be implemented in any method. Recently this was implemented for the generalized van Vleck perturbation approach.⁹³ Szalay⁹⁴ independently formulated similar approach based on Lagrange multipliers for the FSMRCC method. Though in principle, Szalay's approach can be used for general model spaces, this was implemented only for complete model spaces.⁹⁵ Λ -FSMRCC method was successfully implemented for the dipole moment⁹⁶ and polarizability⁹⁷ of the doublet radicals as well as excited states⁹⁸ of molecules. The initial implementation was within singles and doubles (Λ -FSMRCCSD) approximation. Response theory for molecular properties has been pursued by Jorgensen et al. in LR-CC formalism.⁹⁹ Theory for analytic energy derivatives in EOMCC method was proposed by Stanton¹⁰⁰ and implemented by Stanton and Gauss.^{101,102} Nooijen and co-workers implemented gradients in STEOMCC,^{103,104} using Lagrange multipliers. Analytic gradients for SF-EOMCC models at the singles and doubles level has also been proposed recently.¹⁰⁵

However, to improve the accuracy of the molecular properties of the outer valence as well as some of the inner valence states, it is important to include the effects of triples. However, inclusion of full triples is computationally expensive. This limits the applicability of the method to small molecules or to moderate basis sets. Hence, partial inclusion of the triples is more practicable, and this has been implemented in this work. Since triples are added on the basis of perturbative order, it does not guarantee that inclusion of triples will improve molecular properties toward the Full CI (FCI), due to oscillatory nature of the perturbation series. Analytical derivatives for CCSD with various levels of triples excitations has been analyzed long ago.^{106,107} Gauss et al.¹⁰⁸ implemented analytical gradients for the CCSDT model. Recently, parallel calculation of CCSD(T) has been achieved for analytic first and second derivatives.¹⁰⁹ In the context of SRCC, the importance of triples to the dipole moment has also been analyzed.¹¹⁰ Triples excitation in the linear response CC method for excited-state properties was studied iteratively.¹¹¹

In this paper, we present the first implementation of partial triples corrections to the response properties for Lagrange-based formulation within Λ -FSMRCC for the first-order electric property. We have implemented the terms coming from triples whose contribution is at the fourth order in energy and at least up to third order in dipole moment. In Section II, we start with a brief review of the FSMRCC method and the Lagrange approach for energy derivatives. Section III deals with the perturbative analysis of triples amplitude in Λ -FSMRCC for energy and dipole moments. We discuss the results in Section IV.

II. REVIEW OF FSMRCC

The FSMRCC theory^{38,40–42,44} and the Lagrangian formulation within FSMRCC have been described in detail in various articles.^{92,96,97} However, for the completeness of the paper, we briefly discuss the FSMRCC theory here. The FSMRCC method is based on the concept of a common vacuum. We choose an N -electron restricted Hartree–Fock (RHF) as a vacuum. With respect to this vacuum, holes and particles are defined, which are further divided into active and inactive space. Thus, a general model space contains m active particles and n active holes. The model space function can be written as

$$|\Psi_{(0)\mu}^{(m,n)}\rangle = \sum_i C_{\mu i}^{(m,n)} |\Phi_i^{(m,n)}\rangle \quad (1)$$

where, $C_{\mu i}^{(m,n)}$ is the model space coefficient. The correlated wave function for the μ^{th} state can be written as

$$|\Psi_{\mu}^{(m,n)}\rangle = \Omega |\Psi_{(0)\mu}^{(m,n)}\rangle \quad (2)$$

The universal wave operator Ω is such that the states generated by its action on the reference space satisfy the Bloch equation. The wave operator is defined as

$$\Omega = \{e^{\tilde{T}^{(m,n)}}\} \quad (3)$$

The curly bracket denotes normal ordering of the operators within it.¹¹² The cluster operator $\tilde{T}^{(m,n)}$ can be expressed as

$$\tilde{T}^{(m,n)} = \sum_{k=0}^m \sum_{l=0}^n T^{(k,l)} \quad (4)$$

$T^{(k,l)}$ is capable of creating holes and particles in addition to destroying specifically k active particles and l active holes. Thus, $\tilde{T}^{(m,n)}$ amplitudes contain all the lower valence amplitudes and give additional flexibility to the theory. For a specific problem of zero active particle and one active hole, we write the Schrodinger equation for the quasidegenerate states as

$$H|\Psi_{\mu}^{(0,1)}\rangle = E_{\mu}|\Psi_{\mu}^{(0,1)}\rangle \text{ which leads to } H\Omega\left(\sum_i C_{\mu i}^{(0,1)}|\Phi_i^{(0,1)}\rangle\right) = E_{\mu}\Omega\left(\sum_i C_{\mu i}^{(0,1)}|\Phi_i^{(0,1)}\rangle\right) \quad (5)$$

Projection operator for model space is defined as

$$P^{(0,1)} = \sum_i |\Phi_i^{(0,1)}\rangle\langle\Phi_i^{(0,1)}| \quad (6)$$

The complementary space operator Q is $1 - P$. The effective Hamiltonian (H_{eff}) is defined commonly through the Bloch

equation:

$$\begin{aligned} P^{(0,1)}(H\Omega - \Omega H_{\text{eff}}^{(0,1)})P^{(0,1)} &= 0 \\ Q^{(0,1)}(H\Omega - \Omega H_{\text{eff}}^{(0,1)})P^{(0,1)} &= 0 \end{aligned} \quad (7)$$

Because of normal ordering, the contractions among different cluster operators within the exponential are not possible. This leads to decoupling of the equations of different sectors. The equations for the cluster amplitudes are solved, starting from the lowest valence sector upward. This is also known as the subsystem embedding (SEC) condition.

Similar to the Lagrange formulation of linear response approach of SRCC, Szalay⁹⁴ developed a response approach for the multi-reference methods. Though this approach can, in principle, be applied for general model space, this has been implemented to complete model spaces. In this approach response of a specific root out of multiple roots of the effective Hamiltonian is targeted. Thus, one has to project a single desired root of the H_{eff} out of various roots for variation. We construct the Lagrangian and minimize the energy expression with the constraint that the MRCC (i.e., Bloch equations) is satisfied for a specific μ^{th} state:

$$\begin{aligned} \mathcal{J} = & \sum_{ij} \tilde{C}_{\mu i}^{(0,1)} (H_{\text{eff}})_{ij}^{(0,1)} C_{j\mu}^{(0,1)} \\ & + \sum_{ji} \Lambda_{ji}^{(0,1)} \langle \phi_j^{(0,1)} | (H\Omega - \Omega H_{\text{eff}}) | \phi_i^{(0,1)} \rangle \\ & + \sum_{\alpha} \sum_i \Lambda_{\alpha i}^{(0,1)} \langle \phi_{\alpha}^{(0,1)} | (H\Omega - \Omega H_{\text{eff}}) | \phi_i^{(0,1)} \rangle \\ & + \sum_{ji} \Lambda_{ji}^{(0,0)} \langle \phi_j^{(0,0)} | H\Omega | \phi_i^{(0,0)} \rangle \\ & + \sum_{\alpha} \sum_i \Lambda_{\alpha i}^{(0,0)} \langle \phi_{\alpha}^{(0,0)} | H\Omega | \phi_i^{(0,0)} \rangle - E_{\mu} (\sum_{ij} \tilde{C}_{\mu i}^{(0,1)} C_{j\mu}^{(0,1)} - 1) \end{aligned} \quad (8)$$

Where $\phi_i^{(0,1)}$, $\phi_j^{(0,1)}$, $\phi_i^{(0,0)}$, and $\phi_j^{(0,0)}$ are the functions in P space, $\phi_{\alpha}^{(0,1)}$ and $\phi_{\alpha}^{(0,0)}$ are functions in Q space, $\Lambda_{ji}^{(0,1)}$ and $\Lambda_{ji}^{(0,0)}$ are the Lagrange multipliers defined within P space for the (0,1) and (0,0) sectors, respectively. Similarly, $\Lambda_{\alpha i}^{(0,1)}$ and $\Lambda_{\alpha i}^{(0,0)}$ are the Lagrange multipliers from P to Q space for the (0,1) and (0,0) sectors, respectively. However, in case of complete model space (CMS), effective Hamiltonian has an explicit expression in terms of cluster operators, as a result of which the closed part of the Lagrangian multipliers vanishes. Thus, the second and fourth terms of eq 8 vanish, simplifying Lagrangian to

$$\begin{aligned} \mathcal{J} = & \sum_{ij} \tilde{C}_{\mu i}^{(0,1)} (H_{\text{eff}})_{ij}^{(0,1)} C_{j\mu}^{(0,1)} \\ & + \sum_{\alpha} \sum_i \Lambda_{\alpha i}^{(0,1)} \langle \phi_{\alpha}^{(0,1)} | (H\Omega - \Omega H_{\text{eff}}) | \phi_i^{(0,1)} \rangle \\ & + \sum_{\alpha} \sum_i \Lambda_{\alpha i}^{(0,0)} \langle \phi_{\alpha}^{(0,0)} | H\Omega | \phi_i^{(0,0)} \rangle \\ & - E_{\mu} (\sum_{ij} \tilde{C}_{\mu i}^{(0,1)} C_{j\mu}^{(0,1)} - 1) \end{aligned} \quad (9)$$

Differentiation of eq 9 with respect to Λ results in the expression for cluster amplitudes, i.e., the Bloch equation. Differentiation of eq 9 with respect to the T amplitudes leads to equation for Lagrange multipliers. It is seen that the equation for cluster amplitudes is decoupled from the Λ amplitude equation. The Λ equations are however coupled with those of the cluster amplitudes T . In the presence of the external field, the Lagrangian and the parameters $H_{\text{eff}}, C, \tilde{C}, E, \Omega$, and Λ become perturbation dependent. The

differentiation of the Lagrangian with respect to unperturbed cluster amplitude leads to equation for the Lagrangian multipliers. Similarly differentiation of the Lagrangian with respect to unperturbed Lagrange multipliers leads to equation for cluster amplitudes. Cluster amplitudes follow $(2n + 1)$ rule, whereas Lagrange multipliers satisfy $(2n + 2)$ rule. Thus with the help of first derivative of cluster amplitudes and Lagrange multipliers, one can obtain energy derivatives up to second order, i.e., polarizability. Lagrangian for the first- and second-order properties for one valence hole are presented in refs 94 and 95 under singles and doubles approximation. Along similar line, the one valence particle problem can be solved.

III. IMPLEMENTATION OF THE PARTIAL TRIPLES IN Λ -FSMRCC METHOD

In this section, we will present the contribution of triples to the dipole moment, whose origin is beyond the singles and doubles approximations in FSMRCC scheme. Here we will discuss the first implementation of noniterative triples in T and Λ amplitudes to the dipole moment in FSMRCC response. The triples amplitudes are generated as and when used. Since there are several schemes for the inclusion of triples in the literature for SRCC, first we will discuss the specific scheme implemented in this paper for (0,0) sector. The approach implemented here uses canonical orbitals, and the orbitals are not allowed to change with the perturbation, and hence this approach is a nonrelaxed approach. We solve $T_1^{(0,0)}$ and $T_2^{(0,0)}$ amplitudes excluding $VT_3^{(0,0)}$ in a completely iterative manner, which is CCSD approximation. Using these amplitudes of $T_1^{(0,0)}$ and $T_2^{(0,0)}$, $T_3^{(0,0)}$ amplitude is calculated noniteratively from $VT_2^{(0,0)}$ and $VT_2^{(0,0)}T_2^{(0,0)}$. The $T_1^{(0,0)}$ and $T_2^{(0,0)}$ amplitudes are solved iteratively, including the term $VT_3^{(0,0)}$. The inclusion of $VT_3^{(0,0)}$ term in singles and doubles amplitude equations updates the CCSD equations. Even though $VT_3^{(0,0)}$ term is third order, considering the term in a $T_3^{(0,0)}$ equation will make the method iterative. Hence, this term is not included in this scheme. The term $VT_2^{(0,0)}$ contributes at the second order and $VT_2^{(0,0)}T_2^{(0,0)}$ contributes at the third order in perturbation.

In the implementation of the MRCCSD(T^*)/CCSD(T^*) approximation, we construct an intermediate operator \bar{H} given by ($\bar{H} = \exp(-T^{(0,0)})H\exp(T^{(0,0)})$) and truncate up to one(\bar{F}), two(\bar{V}) and three body(\bar{W}) parts. For the construction of \bar{H} , we use CCSD approximation without including the amplitudes of triples, i.e., $T_3^{(0,0)}$. H_{eff} under this approximation is

$$\begin{aligned} H_{\text{eff}} = & P^{(0,1)}(\bar{F} + \bar{F}T_1^{(0,1)} + \bar{V}T_2^{(0,1)} \\ & + \bar{F}T_2^{(0,1)} + \bar{V}T_3^{(0,1)})P^{(0,1)} \end{aligned} \quad (10)$$

The Fock-space Bloch equations for the $T_1^{(0,1)}$, $T_2^{(0,1)}$, and $T_3^{(0,1)}$ amplitudes are as below:

$$\begin{aligned} Q_1^{(0,1)}(\bar{F} + \bar{F}T_1^{(0,1)} + \bar{V}T_2^{(0,1)} \\ + \bar{F}T_2^{(0,1)} + \bar{V}T_3^{(0,1)} - T_1^{(0,1)}H_{\text{eff}})P^{(0,1)} = 0 \end{aligned} \quad (11)$$

$$\begin{aligned} Q_2^{(0,1)}(\bar{V} + \bar{F}T_2^{(0,1)} + \bar{V}T_1^{(0,1)} + \bar{V}T_2^{(0,1)} \\ + \bar{W}T_2^{(0,1)} + \bar{V}T_3^{(0,1)} + \bar{F}T_3^{(0,1)} - T_2^{(0,1)}H_{\text{eff}})P^{(0,1)} = 0 \end{aligned} \quad (12)$$

$$\begin{aligned} Q_3^{(0,1)}(\bar{W} + \bar{W}T_2^{(0,1)} + \bar{F}T_3^{(0,1)} + \bar{V}T_2^{(0,1)} - T_3^{(0,1)}H_{\text{eff}})P^{(0,1)} \\ = 0 \end{aligned} \quad (13)$$

Table 1. Dipole Moments of $^2\Pi$ OH Radical^a

basis	Λ -FSMRCCSD(T*)	Λ -FSMRCCSD	EOMCCSD(unrelaxed) ^b	full CI ^c
cc-pVDZ	0.682	0.634	0.639	0.663
cc-pVTZ	0.682	0.645	—	—
cc-pVQZ	0.684	0.645	—	—

^a Results in au and $R_{\text{eq}} = 1.85104 a_0$. ^b See ref 96. ^c See ref 113.

It can be seen that $\bar{V}T_3^{(0,1)}$ is the only term contributing to the singles and doubles amplitude equation along with H_{eff} . It is easy to see that \bar{W} cannot contribute to H_{eff} . The eqs 11 and 12 are first solved fully excluding the terms which involve $T_3^{(0,1)}$ amplitude, which is the CCSD approximation. Using these amplitudes eq 13 is solved noniteratively. In eq 13, we want to be accurate up to third order. Hence we include in the term $T_3^{(0,1)}H_{\text{eff}}$ only $T_3^{(0,1)}\bar{F}$. After solving $T_3^{(0,1)}$, we again solve the eqs 11 and 12 iteratively. Here the effect of $T_3^{(0,1)}$ appears via $\bar{V}T_3^{(0,1)}$ and $\bar{F}T_3^{(0,1)}$.

We now consider the triples correction to the Λ amplitudes and then to the overall dipole moment. The Λ equations are like the conjugates of the T amplitude equations, and hence the terms in T equations appear in Λ equations also. It should be mentioned here that unlike in T amplitude equations, we first solve for the (0,1) sector and then for the (0,0) sector due to reverse decoupling in Λ equations.

First, the Λ amplitudes in singles and doubles approximation are solved iteratively for both (0,1) and (0,0) sector. With these Λ amplitudes the Lagrangian for triples is constructed. During the construction, the singles and doubles (SD) terms remain as such. The Lagrangian with the triples correction is given by

$$\begin{aligned} \mathcal{L} = & SD + \bar{V}T_3^{(0,1)}\bar{C}\bar{C} + \Lambda_3^{(0,1)}\bar{V}T_2^{(0,1)} + \Lambda_3^{(0,1)}\bar{W}T_2^{(0,1)} \\ & + \Lambda_3^{(0,1)}\bar{F}T_3^{(0,1)} + \Lambda_2^{(0,1)}\bar{V}T_3^{(0,1)} + \Lambda_2^{(0,1)}\bar{F}T_3^{(0,1)} \\ & - \Lambda_2^{(0,1)}T_2^{(0,1)}(\bar{V}T_3^{(0,1)}) + \Lambda_1^{(0,1)}\bar{V}T_3^{(0,1)} + \Lambda_3^{(0,0)}VT_2^{(0,0)} \\ & + \Lambda_2^{(0,0)}VT_3^{(0,0)} + \Lambda_1^{(0,0)}VT_3^{(0,0)} + \Lambda_3^{(0,0)}FT_3^{(0,0)} \\ & + \Lambda_3^{(0,1)}\bar{W}T_2^{(0,0)} + \Lambda_3^{(0,1)}VT_3^{(0,0)} + \Lambda_3^{(0,1)}VT_2^{(0,0)} \end{aligned} \quad (14)$$

The \bar{C} and C are left and right eigen vectors of the H_{eff} . The Lagrangian in eq 14 is differentiated with respect to $T_3^{(0,1)}$ to get the equation for $\Lambda_3^{(0,1)}$. The equation defining the $\Lambda_3^{(0,1)}$ amplitude is given in eq 15:

$$\begin{aligned} \langle P^{(0,1)} | \bar{V}\bar{C}\bar{C} + \Lambda_3^{(0,1)}\bar{F} - \Lambda_2^{(0,1)}T_2^{(0,1)}\bar{V} + \Lambda_1^{(0,1)}\bar{V} \\ + \Lambda_2^{(0,1)}\bar{V} | Q^{(0,1)} \rangle = 0 \end{aligned} \quad (15)$$

The Lagrangian in eq 14 is differentiated with respect to $T_2^{(0,1)}$ to get the equation for $\Lambda_2^{(0,1)}$. The $\Lambda_2^{(0,1)}$ equation is given by

$$SD + \langle P^{(0,1)} | \Lambda_3^{(0,1)}\bar{V} + \Lambda_3^{(0,1)}\bar{W} | Q^{(0,1)} \rangle = 0 \quad (16)$$

The eq 15 is solved noniteratively to obtain $\Lambda_3^{(0,1)}$. The connected terms in $\Lambda_3^{(0,1)}$ amplitude equation are considered in this approximation. Thus, $\Lambda_3^{(0,1)}$ amplitude is obtained from the terms $\bar{V}\bar{C}\bar{C}$, $\Lambda_2^{(0,1)}\bar{V}$, $\Lambda_1^{(0,1)}\bar{V}$, and $\Lambda_2^{(0,1)}T_2^{(0,1)}\bar{V}$. These contribute at the first, second, and third orders, respectively. The $\Lambda_3^{(0,1)}$ amplitude is also obtained from the second order \bar{F} containing term $\Lambda_3^{(0,1)}\bar{F}$. After obtaining $\Lambda_3^{(0,1)}$, its effect on

$\Lambda_2^{(0,1)}$ appears through the third-order terms $\Lambda_3^{(0,1)}\bar{V}$ and $\Lambda_3^{(0,1)}\bar{W}$. The equation for $\Lambda_2^{(0,1)}$ amplitude, eq 16 is solved fully by taking into account the $\Lambda_3^{(0,1)}$ terms calculated above.

For solving (0,0) sector $\Lambda_3^{(0,0)}$ amplitude is obtained first. Here too the equation for $\Lambda_3^{(0,0)}$ is obtained by differentiating the Lagrangian in eq 14 with respect to $T_3^{(0,0)}$. The terms appear after the differentiation with respect to $T_3^{(0,0)}$ are given in eq 17:

$$\begin{aligned} \langle P^{(0,0)} | \Lambda_3^{(0,0)}F + \Lambda_2^{(0,0)}V + \Lambda_1^{(0,0)}V + \Lambda_3^{(0,1)}V | Q^{(0,0)} \rangle \\ = 0 \end{aligned} \quad (17)$$

The equation for $\Lambda_2^{(0,0)}$ is obtained from differentiating eq 14 with respect to $T_2^{(0,0)}$. The $\Lambda_2^{(0,0)}$ equation with triples correction is given by

$$\begin{aligned} SD + \langle P^{(0,0)} | \Lambda_3^{(0,0)}V + \Lambda_3^{(0,1)}V + \Lambda_3^{(0,0)}VT_2^{(0,0)} | Q^{(0,0)} \rangle \\ = 0 \end{aligned} \quad (18)$$

The eq 17 is solved noniteratively to obtain $\Lambda_3^{(0,0)}$. $\Lambda_3^{(0,0)}$'s are obtained by taking the direct contribution from the second-order term $\Lambda_2^{(0,0)}V$ and the third-order terms $\Lambda_1^{(0,0)}V$ and $\Lambda_3^{(0,1)}V$. Also F containing term $\Lambda_3^{(0,0)}F$ contributes to $\Lambda_3^{(0,0)}$ equation at second order. It should be noted that due to reverse decoupling $\Lambda^{(0,1)}$ involving terms $\Lambda_3^{(0,1)}VT_2^{(0,0)}$ and $\Lambda_3^{(0,1)}VT_3^{(0,0)}$ appears in $\Lambda^{(0,0)}$. After obtaining $\Lambda_3^{(0,0)}$, its effect on $\Lambda_2^{(0,0)}$ equation is incorporated through the third-order terms $\Lambda_3^{(0,0)}V$, $\Lambda_3^{(0,1)}V$, and $\Lambda_3^{(0,0)}VT_2^{(0,0)}$. The eq 18 is solved fully by taking into account of $\Lambda_3^{(0,0)}$ terms calculated above. Finally, the triples contribution to $E^{(1)}$ is given in eq 19, where \hat{O} is the explicit derivative of Hamiltonian with respect to external field:

$$E_{\text{triples}}^{(1)} = \Lambda_2^{(0,1)}\hat{O}T_3^{(0,1)} + \Lambda_2^{(0,0)}\hat{O}T_3^{(0,0)} \quad (19)$$

These triples corrected Λ and T amplitudes are used for the evaluation of dipole moments in (0,0) and (0,1) sectors. The term $VT_2^{(0,0)}T_2^{(0,0)}$ in $T_3^{(0,0)}$ equation will thus have a higher effect on the dipole moment, while the other triples correcting terms will affect the dipole moment at the third order. The third-order terms which appear in the dipole moment are $\Lambda_2^{(0,1)}\hat{O}T_3^{(0,1)}$ and $\Lambda_2^{(0,0)}\hat{O}T_3^{(0,0)}$. Hence the final dipole moment is corrected at least up to third order in triples.

IV. RESULTS AND DISCUSSION

We have implemented the contribution of triples partially to the FSMRCC singles and doubles scheme (FSMRCCSD(T*)). To test our code we chose small systems as a case study. We present our results and discussion on them in this section. The code is tested against the nonrelaxed finite field approach. The systems studied are $\dot{\text{O}}\text{H}$, $\dot{\text{O}}\text{O}\text{H}$, $\text{HCO}\dot{\text{O}}$, $\dot{\text{C}}\text{N}$, and $\dot{\text{C}}\text{H}$.

A. OH Radical. We report the dipole moment of hydroxy radical at the equilibrium geometry in Table 1. We start with the closed-shell configuration of OH^- anion as a vacuum. The highest occupied molecular orbital (HOMO) of OH^- is two-

Table 2. Dipole Moments of $^2\Sigma^+$ State of CN Radical^a

basis	ROHF ^b		Λ -FSMRCC	
	CCSD	CCSD(T)	CCSD	CCSD(T*)
cc-pVDZ	0.522	0.476	0.427 (0.437) ^c	0.497 (0.489) ^c
aug-cc-pVDZ			0.510	0.558
CBS limit ^d			0.559 \pm 0.001	
exp ^e			0.57 \pm 0.03	

^a Results in au and $R_{\text{eq}} = 2.21512 a_0$. ^b Results obtained from ACES II package. ^c Relaxed finite-field values. ^d See ref 114. ^e See ref 116.

fold degenerate in nature. The degenerate HOMO's are chosen as active holes of the Fock space (0,1) sector. The removal of an electron from one of these HOMO's lead to degenerate doublet $^2\Pi$ of the hydroxy radical. In Table 1 we report the dipole moment of hydroxy radical in cc-pVXZ (X = D, T, Q) basis. The calculated FCI and available EOMCCSD dipole moments in cc-pVDZ basis are also presented. The Λ -FSMRCCSD values show that the dipole moment is converged from cc-pVDZ to cc-pVQZ. Whereas, the Λ -FSMRCCSD(T*) produces a marginal change in dipole moment. It is observed that the Λ -FSMRCCSD dipole moment in cc-pVDZ basis is 0.634 au, whereas the Λ -FSMRCCSD(T*) increases the dipole moment (0.682 au) toward the FCI value of 0.663 au.¹¹³ Though triples exceeds the FCI dipole moment, the qualitative trend toward FCI dipole moment is obtained. With the higher order triples it may improve further.

B. CN Radical. The dipole moments of the CN radical are presented in the Table 2. We start with the cyanide anion, which is closed shell with the ground-state geometry $^2\Sigma^+$. Removal of an electron from the cyanide anion gives CN radical. The studies are carried out with one active hole. Since the dipole moment of the CN radical is important in astrophysics, there are various theoretical calculations^{114,115} to achieve the experimental accuracy. In Table 2 we report the dipole moment obtained using our method in cc-pVDZ and aug-cc-pVDZ basis sets. For cc-pVDZ basis we also report the finite field dipole moment values using restricted open-shell Hartree–Fock (ROHF)-CC and FSMRCC within singles and doubles approximation as well as with partial triples. The values presented in parentheses denote the finite field FSMRCC results. Observation of the various levels of theory^{114,115} says that it is necessary to have augmented basis sets for the dipole moment calculations of CN radical, which is clearly reflected in the Table 2. It has been observed so far that only beyond the double- ζ with augmentation, the dipole moment close to CBS limit (0.559 \pm 0.001 au)¹¹⁴ and experimental (0.57 \pm 0.03 au)¹¹⁶ value is attained. In our method, as we go from cc-pVDZ basis to aug-cc-pVDZ basis, Λ -FSMRCCSD gives dipole moment values of 0.427 and 0.510 au, respectively. Thus, with augmented basis at the CCSD level is close to the reported CBS limit as well as experimental dipole moment. The inclusion of triples improves the dipole moment values to 0.497 au for cc-pVDZ basis and 0.558 au for aug-cc-pVDZ basis, respectively. It can be seen that the qualitative trend remains the same in both the basis sets, i.e., triples correction increases the dipole moment values. However, in ROHF-CC approach the trend is opposite to that of Λ -FSMRCC. ROHF-CC results are obtained using finite field approach which includes relaxation effects. To test the effect of relaxation we have done a finite field relaxed FSMRCC

calculation. Here too, we get the same trend as we obtained from the analytic nonrelaxed approach. Thus the difference in trends of dipole moment in Λ -FSMRCC and ROHF-CC may arise due to combination of the way triples are included and the treatment of dynamic correlation.

C. OOH and HCOO Radicals. The dipole moments for the nonlinear molecules, such as hydroperoxy and formyloxy radical, at the equilibrium geometry were studied using the double- ζ basis set of Huzinaga–Dunning^{117,118} with a set of uncontracted polarized functions. The description of the geometries for these radicals is given in Appendix I. The center of mass coordinates is used, and the molecules are kept along the X,Y direction. The dipole moments for each direction are obtained and presented in Table 3. Since there is no FCI or experimental dipole moment available for these systems, we report the relaxed finite field FSMRCC (FF-FSMRCC) dipole moments. We start with the RHF of hydroperoxide anion as vacuum. The electronic configuration of RHF of hydroperoxide anion is [core], $3a_1^2, 4a_1^2, 5a_1^2, 1a_2^2, 6a_1^2, 7a_1^2, 2a_2^2$.

Removal of an electron from one of the two highest occupied orbitals results in near-degenerate states (2A_2 and 2A_1) of hydroperoxy radicals. The dipole moments of the radical along two orthogonal directions (X and Y) have been presented in Table 3. We also report the FF-FSMRCC calculations for the system. In this case, the Λ -FSMRCCSD(T*) predicts the lower dipole moment than one obtained from the FF-FSMRCCSD(T*) method.

The dipole moments of the first two low-lying near-degenerate states of the formyloxy radical at the equilibrium geometry are given in Table 3. We start with the RHF of formate anion as vacuum. Removal of an electron from the formate anion results in formyloxy radical, the near degenerate low-lying states of which have the electronic configuration: [core], $3a_1^2, 2b_2^2, 4a_1^2, 5a_1^2, 3b_2^2, 1b_1^2, 1a_2^2, 6a_1^2, 4b_2^1$ and [core], $3a_1^2, 2b_2^2, 4a_1^2, 5a_1^2, 3b_2^2, 1b_1^2, 1a_2^2, 6a_1^1, 4b_2^2$. The dipole moments along the H–C bond axis for these states, denoted by 2B_2 and 2A_1 have been reported. The EOMCC result¹⁰¹ for the ground state has also been reported. We have also mentioned the finite field dipole moment obtained by the FF-FSMRCCSD(T*) in Table 3, which stays close to the dipole moment obtained from the Λ -FSMRCCSD(T*) method.

D. CH Radical. The CH radical can be considered as the electron attached state of the corresponding cation CH^+ . The RHF configuration of CH^+ , $1\sigma^2 2\sigma^2 3\sigma^2$ is chosen as a vacuum. The degenerate lowest unoccupied molecular orbitals (LUMO's) are chosen as active particles. For CH^+ we report the dipole moment at the equilibrium as well as at the stretched geometry, i.e., at 1.5 R_e . Table 4 reports the results for the CH radical in cc-pVDZ¹¹⁹ and Sadlej¹²⁰ basis along the direction of molecular axis. We compare the dipole moment obtained from cc-pVDZ basis with the FCI dipole moment and the dipole moment obtained from Sadlej basis with experimental¹²¹ value. At the equilibrium geometry the dipole moment value is reduced in Λ -FSMRCCSD as well as in Λ -FSMRCCSD(T*) as we go from cc-pVDZ to Sadlej basis. However, at the stretched geometry, the dipole moment is increased with the basis set. In cc-pVDZ basis, at the equilibrium geometry the Λ -FSMRCCSD dipole moment value is 0.582 au, which is reduced by the triples correction (0.575 au). The Λ -FSMRCCSD(T*) dipole moment (0.575 au) is closer to the FCI (0.548 au) value. At the stretched geometry the Λ -FSMRCCSD gives 0.100 au, the inclusion of the triples reduces it to 0.061 au, which is approaching toward the FCI value of 0.074 au. This

Table 3. Dipole Moments of OOH and HCOO Radical^a

state	direction	Λ -FSMRCCSD(T*)	Λ -FSMRCCSD	FF-FSMRCCSD(T*)	EOMCCSD ^b
OOH					
² A ₂	X	-0.588	-0.557	-0.571	
	Y	-0.713	-0.669	-0.692	
total		0.924	0.870	0.897	
¹ A ₂	X	-0.402	-0.369	-0.387	
	Y	-0.717	-0.676	-0.694	
total		0.822	0.770	0.795	
HCOO					
² B ₂	Y	0.965	0.909	0.979	1.004
² A ₁	Y	0.835	0.786	0.842	—

^b See ref 101. ^a Results in au.

Table 4. Dipole Moments of CH Radical^a

basis	Λ -FSMRCCSD(T*)	Λ -FSMRCCSD	full CI
cc-pVDZ <i>R</i> _{eq}	0.575	0.582	0.548
Sadlej <i>R</i> _{eq}	0.547	0.540	—
exp ^b			0.57 ± 0.023
cc-pVDZ <i>R</i> _{dis}	0.061	0.100	0.074
Sadlej <i>R</i> _{dis}	0.084	0.111	—

^a Results in au, *R*_{eq} = 2.11648 a₀, and *R*_{dis} = 3.1660 a₀. ^b See ref 121.

Table 5. Dipole Moments of PO Radical^a

basis	ROHF ^b		Λ -FSMRCC	
	CCSD	CCSD(T)	CCSD	CCSD(T*)
cc-pVDZ	0.777	0.726	0.708	0.750
exp ^c				0.740 ± 0.028

^a Results in au and *R*_{eq} = 2.78357 a₀. ^b Results obtained from the ACES II package ^c See ref 122.

emphasizes the importance of inclusion of the triples for the calculation of dipole moment at the stretched geometry. Similar trend is observed for the Sadlej basis too. However, even at the equilibrium geometry with the inclusion of partial triples the dipole moment value approaches toward FCI. This shows the importance of the triples even at the equilibrium geometry.

E. PO Radical. The dipole moment of PO radical, which is difficult to predict by the single reference method, has been studied using FSMCCSD and FSMRCCSD(T*). The RHF configuration of PO⁻ has been taken as the vacuum. The calculations are carried out with one active hole. The dipole moment value of PO radical obtained using cc-pVDZ basis in FSMRCCSD is 0.708 au and FSMRCCSD(T*) is 0.750 au. The dipole moment obtained from the FSMRCCSD(T*) method, as can be seen from the Table 5, is slightly overestimated. However, inclusion of triples improves the accuracy toward the experimental value of 1.88 ± 0.07 debye (0.740 ± 0.028 au). The finite field relaxed ROHF-CCSD and ROHF-CCSD(T) are performed in the same basis using the ACES-II package.¹²³ Finite field calculations using large basis sets are reported for this radical at ROHF-CCSD(T) level by Urban et al.¹¹⁵ The opposite trend

in the inclusion of triples is observed for ROHF-based CCSD and CCSD(T). This difference in trends for dipole moment on inclusion of triples in Λ -FSMRCC and ROHF-CC could be due to a combination of reasons like the way triples are included and the treatment of dynamic correlation.

V. CONCLUSIONS

In this paper we presented the implementation and the results for the recently developed Lagrange-based Fock-space multi-reference coupled cluster response approach with the inclusion of partial triples for electric properties of the doublet radicals. The results for the OH and CH indicate that Λ -FSMRCCSD(T*) performs better than Λ -FSMRCCSD and tends toward FCI. In particular, when the dipole moments of Λ -FSMRCCSD and Λ -FSMRCCSD(T*) are compared at the 1.5*R*_e for the CH radical, we can observe that the inclusion of triples leads to more accurate results than that of the Λ -FSMRCCSD results. At stretched geometries, where the multireference description is required, inclusion of the triples provides better results. From the dipole moment of OH radical using cc-pVDZ, TZ, and QZ basis sets, it is observed that at the Λ -FSMRCCSD level dipole moment saturates at 0.645 au, whereas with the inclusion of triples it is 0.684 au, which tends to the FCI dipole moment of 0.663 au. Though it is slightly overestimated, compared to the FCI it gives qualitatively correct trend. Also, the nonrelaxed EOMCCSD shows a dipole moment of 0.639 au which is closer to the Λ -FSMRCCSD value. The results of the analytic Λ -FSMRCCSD(T*) are compared with the finite field dipole moments for OOH and HCOO molecules. In both the cases, it is observed that the analytic Λ -FSMRCCSD(T*) shows a qualitatively correct trend, as does the finite field dipole moment. However, it should be mentioned here that the finite field method has explicit relaxation through the orbital rotation, whereas the analytic method implemented does not include the explicit relaxation effects. The calculations are performed for CN radical using cc-pVDZ and aug-cc-pVDZ basis sets. Augmented basis set helps to get the results closer to the basis set limit. Inclusion of the noniterative triples improves dipole moment by about 9%. The inclusion of the triples indicates the dipole moment closer to the experimental as well as basis set limit value. Another radical where we have analyzed the importance of triples excitation is PO. We observe that the triples excitation

Table 6. Geometries in au

molecule	atom	X	Y	Z
ÖOH	H	-1.60075	-1.66668	0.00000
	O	1.27888	-0.01807	0.00000
	O	-1.17802	0.12308	0.00000
HCOÖ	H	0.00000	2.96725	0.00000
	C	0.00000	0.88855	0.00000
	O ₁	-1.98007	-0.42700	0.00000
	O ₂	1.98007	-0.42700	0.00000

improves the results for cc-pVDZ basis. ROHF-based CCSD and CCSD(T) calculations are performed to analyze the way triples improve the dipole moment for CN and PO radicals. It has been observed that the way the triples contributes to the dipole moment is opposite to that of the FSMRCC method. This can be due to the different way the triples are taken in FSMRCCSD-(T*) method and the different treatment of the dynamic correlation. Thus, all the results emphasize the importance of triples for the accurate calculation of the dipole moment for the doublet radicals.

APPENDIX I

Geometries in au are summarized in Table 6.

AUTHOR INFORMATION

Corresponding Author

*E-mail: s.pal@ncl.res.in.

ACKNOWLEDGMENT

The authors acknowledge the facilities of the Center of Excellence in Scientific Computing at NCL. One of the authors, L.R., would like to thank the Council of Scientific and Industrial Research (CSIR) for a Senior Research Fellowship. S.P. acknowledges grant from DST J. C. Bose Fellowship project towards completion of the work. N.V. would like to thank the Department of Science and Technology, India, for financial support.

REFERENCES

- (1) Cizek, J. *Adv. Chem. Phys.* **1969**, *14*, 35.
- (2) Cizek, J. *J. Chem. Phys.* **1966**, *45*, 4256.
- (3) Paldus, J.; Cizek, J.; Shavitt, I. *Phys. Rev. A: At., Mol., Opt. Phys.* **1972**, *5*, 50.
- (4) Bartlett, R. J. *Annu. Rev. Phys. Chem.* **1981**, *32*, 359.
- (5) Coester, F. *Nucl. Phys.* **1958**, *7*, 421.
- (6) Kummel, H. *Nucl. Phys.* **1960**, *17*, 477.
- (7) Bartlett, R. J. In *Modern Electronic Structure Theory*; Yarkony, D. R., Ed.; World Scientific: Singapore, 1995; Vol. 2, p 1047.
- (8) Bartlett, R. J. In *Geometrical Derivatives of Energy Surfaces and Molecular Properties*; Jorgensen, P., Simons, J., Eds.; Reidel: Dordrecht, The Netherlands, 1986, p 3561.
- (9) Adamowicz, L.; Laidig, W. D.; Bartlett, R. J. *Int. J. Quantum Chem.* **1984**, *18*, 245.
- (10) Scheiner, A. C.; Scuseria, G.; Lee, T. J.; Rice, J.; Schaefer, H. F., III *J. Chem. Phys.* **1987**, *87*, 5361.
- (11) Wang, F.; Gauss, J. *J. Chem. Phys.* **2008**, *129*, 174110.
- (12) Kallay, M.; Gauss, J.; Szalay, P. G. *J. Chem. Phys.* **2003**, *119*, 2991.
- (13) Monkhorst, H. *Int. J. Quantum Chem.* **1977**, *S11*, 421.
- (14) Salter, E. A.; Trucks, G.; Bartlett, R. J. *J. Chem. Phys.* **1989**, *90*, 1752.
- (15) Helgaker, T.; Jorgensen, P. *Adv. Quantum Chem.* **1988**, *19*, 183.
- (16) Christiansen, O.; Koch, A.; Jorgensen, P. *Chem. Phys. Lett.* **1995**, *243*, 409.
- (17) Koch, H.; Christiansen, O.; Jorgensen, P.; de Meras, A. S.; Helgaker, T. *J. Chem. Phys.* **1997**, *106*, 1808.
- (18) Medved, M.; Urban, M.; Noga, J. *Theor. Chem. Acc.* **1997**, *98*, 75.
- (19) Scuseria, G. E.; Schaefer, H. F., III *Chem. Phys. Lett.* **1988**, *146*, 23.
- (20) Noga, J.; Bartlett, R. J. *J. Chem. Phys.* **1987**, *86*, 7041.
- (21) Kucharski, S. A.; Bartlett, R. J. *J. Chem. Phys.* **1998**, *108*, 5243.
- (22) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479.
- (23) Bartlett, R. J.; Watts, J. D.; Kucharski, S. A.; Noga, J. *Chem. Phys. Lett.* **1990**, *165*, 513.
- (24) Urban, M.; Noga, J.; Cole, S. J.; Bartlett, R. J. *J. Chem. Phys.* **1985**, *83*, 4041.
- (25) Bomble, Y. J.; Stanton, J. F.; Kallay, M.; Gauss, J. *J. Chem. Phys.* **2005**, *123*, 054101.
- (26) Medved, M.; Urban, M.; Kello, V.; Diercksen, G. H. F. *J. Mol. Struct. Theochem* **2001**, *547*, 219.
- (27) Piecuch, P.; Adamowicz, L. *J. Chem. Phys.* **1994**, *100*, 5857.
- (28) Oliphant, N.; Adamowicz, L. *J. Chem. Phys.* **1991**, *94*, 1229.
- (29) Oliphant, N.; Adamowicz, L. *J. Chem. Phys.* **1992**, *96*, 3739.
- (30) Ghose, K. B.; Piecuch, P.; Pal, S.; Adamowicz, L. *J. Chem. Phys.* **1996**, *104*, 6582.
- (31) Meissner, L.; Jankowski, K.; Wasilewski, J. *Int. J. Quantum Chem.* **1988**, *34*, 535.
- (32) Hurtubise, V.; Freed, K. F. *Adv. Chem. Phys.* **1993**, *83*, 465.
- (33) Durand, P.; Malrieu, J. P. *Adv. Chem. Phys.* **1987**, *67*, 321.
- (34) Evangelisti, S.; Daudey, J. P.; Malrieu, J. P. *Phys. Rev. A: At., Mol., Opt. Phys.* **1987**, *35*, 4930.
- (35) Jeziorski, B.; Monkhorst, H. J. *Phys. Rev. A: At., Mol., Opt. Phys.* **1981**, *24*, 1668.
- (36) Paldus, J.; Pylypow, L.; Jeziorski, B. In *Many-body methods in quantum chemistry, Lecture notes in chemistry*; Kaldor, U., Ed.; 1989, Vol. 52, p 151.
- (37) Balkova, A.; Kucharski, S. A.; Meissner, L.; Bartlett, R. J. *J. Chem. Phys.* **1991**, *95*, 4311.
- (38) Kutzelnigg, W. *J. Chem. Phys.* **1982**, *77*, 2081.
- (39) Mukherjee, D.; Moitrie, R. K.; Mukhopadhyay, A. *Mol. Phys.* **1975**, *30*, 1861. (1975);
- (40) Mukherjee, D. *Pramana* **1979**, *12*, 203.
- (41) Mukherjee, D.; Pal, S. *Adv. Quantum Chem.* **1989**, *20*, 292.
- (42) Lindgren, L.; Mukherjee, D. *Phys. Rep.* **1987**, *151*, 93.
- (43) Haque, M.; Kaldor, U. *Chem. Phys. Lett.* **1985**, *117*, 347.
- (44) Pal, S.; Rittby, M.; Bartlett, R. J.; Sinha, D.; Mukherjee, D. *J. Chem. Phys.* **1988**, *88*, 4357.
- (45) Chattopadhyay, S.; Mahapatra, U. S.; Datta, B.; Mukherjee, D. *Chem. Phys. Lett.* **2002**, *357*, 426.
- (46) Chattopadhyay, S.; Mahapatra, U. S.; Mukherjee, D. *J. Chem. Phys.* **1999**, *111*, 3820.
- (47) Meissner, L. *J. Chem. Phys.* **1998**, *108*, 9227.
- (48) Meissner, L. *Chem. Phys. Lett.* **1996**, *255*, 244.
- (49) Landau, A.; Eliav, E.; Ishikawa, Y.; Kaldor, U. *J. Chem. Phys.* **2004**, *121*, 6634.
- (50) Eliav, E.; Borschevsky, A.; Shamasundar, K. R.; Pal, S.; Kaldor, U. *Int. J. Quantum Chem.* **2009**, *109*, 2909.
- (51) Geertsen, J.; Rittby, M.; Bartlett, R. J. *Chem. Phys. Lett.* **1989**, *164*, 57.
- (52) Stanton, J. F.; Bartlett, R. J. *J. Chem. Phys.* **1993**, *98*, 7029.
- (53) Bartlett, R. J.; Stanton, J. F. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 1994; Vol. 5, p 65.
- (54) Pieniazek, P. A.; Arnstein, S. A.; Bradforth, S. E.; Krylov, A. I.; Sherrill, C. D. *J. Chem. Phys.* **2007**, *127*, 164110.

- (55) Nooijen, M.; Bartlett, R. J. *J. Chem. Phys.* **1995**, *102*, 3629.
- (56) Kowalski, K.; Piecuch, P. *Chem. Phys. Lett.* **2001**, *347*, 237.
- (57) Wladyslawski, M.; Nooijen, M. In *Low-lying potential energy surfaces*, ACS Symposium Series; Hoffmann, M. R., Dyall, K. G., Eds.; American Chemical Society: Washington, D.C., 2002; Vol. 828, p 65.
- (58) Mukherjee, D.; Mukherjee, P. K. *Chem. Phys.* **1979**, *39*, 325.
- (59) Koch, H.; Jorgensen, P. *J. Chem. Phys.* **1990**, *93*, 3333.
- (60) Koch, H.; Aa.Jensen, H. J.; Jorgensen, P.; Helgaker, T. *J. Chem. Phys.* **1990**, *93*, 3345.
- (61) Nooijen, M.; Bartlett, R. J. *J. Chem. Phys.* **1997**, *106*, 6441.
- (62) Nooijen, M.; Bartlett, R. J. *J. Chem. Phys.* **1997**, *106*, 6449.
- (63) Krylov, A. I. *Chem. Phys. Lett.* **2001**, *338*, 375.
- (64) Nakatsuji, H.; Hirao, K. *J. Chem. Phys.* **1978**, *68*, 2053.
- (65) Ehara, M.; Nakatsuji, H. *J. Chem. Phys.* **1993**, *99*, 1952.
- (66) Piecuch, P.; Kowalski, K.; Pimienta, I. S. O.; Kucharski, S. A. In *Low-lying potential energy surfaces*, ACS Symposium Series; Hoffmann, M. R., Dyall, K. G., Eds.; American Chemical Society: Washington, D.C., 2002; Vol. 828, p 31.
- (67) Piecuch, P.; Kowalski, K.; Pimienta, I. S. O.; Fan, P. D.; Lodriguito, M.; McGuire, M. J.; Kucharski, S. A.; Kus, T.; Musial, M. *Theor. Chem. Acc.* **2004**, *112*, 349.
- (68) Pal, S.; Rittby, M.; Bartlett, R. J. *Chem. Phys. Lett.* **1989**, *160*, 212.
- (69) Val, N.; Ghose, K. B.; Pal, S.; Mukherjee, D. *Chem. Phys. Lett.* **1993**, *209*, 292.
- (70) Val, N.; Pal, S.; Mukherjee, D. *Theor. Chem. Acc.* **1998**, *99*, 100.
- (71) Musial, M.; Bartlett, R. J. *J. Chem. Phys.* **2008**, *129*, 244111.
- (72) Musial, M.; Bartlett, R. J. *J. Chem. Phys.* **2008**, *129*, 044101.
- (73) Kowalski, K.; Piecuch, P. *J. Chem. Phys.* **2000**, *113*, 8490.
- (74) Pieniazek, P. A.; Bradforth, S. E.; Krylov, A. I. *J. Chem. Phys.* **2008**, *129*, 074104.
- (75) Kallay, M.; Gauss, J. *J. Chem. Phys.* **2004**, *121*, 9257.
- (76) Kamiya, M.; Hirata, S. *J. Chem. Phys.* **2006**, *125*, 074111.
- (77) Musial, M.; Kucharski, S. A.; Bartlett, R. J. *J. Chem. Phys.* **2003**, *118*, 1128.
- (78) Kowalski, K.; Piecuch, P. *J. Chem. Phys.* **2002**, *116*, 7411.
- (79) Evangelista, F. A.; Prochnow, E.; Gauss, J.; Schaefer, H. F., III *J. Chem. Phys.* **2007**, *132*, 074107.
- (80) Stanton, J. F.; Gauss, J. *Theor. Chim. Acta.* **1996**, *93*, 303.
- (81) Manohar, P. U.; Stanton, J. F.; Krylov, A. I. *J. Chem. Phys.* **2009**, *131*, 114112.
- (82) Kowalski, K.; Piecuch, P. *J. Phys. Chem.* **2001**, *115*, 643.
- (83) Manohar, P. U.; Krylov, A. I. *J. Chem. Phys.* **2008**, *129*, 194105.
- (84) Kowalski, K.; Piecuch, P. *J. Chem. Phys.* **2001**, *115*, 2966.
- (85) Jorgensen, P.; Simons, J. In *Second Quantization Based Methods in Quantum Chemistry*; Academic Press: New York, 1981.
- (86) Olsen, J.; Jorgensen, P. *J. Chem. Phys.* **1985**, *82*, 3235.
- (87) Pal, S. *Phys. Rev. A: At., Mol., Opt. Phys.* **1989**, *39*, 39.
- (88) Pal, S. *Int. J. Quantum Chem.* **1992**, *41*, 443.
- (89) Ajitha, D.; Pal, S. *J. Chem. Phys.* **2001**, *114*, 3380.
- (90) Ajitha, D.; Val, N.; Pal, S. *J. Chem. Phys.* **1999**, *110*, 2316.
- (91) Shamasundar, K. R.; Pal, S. *J. Chem. Phys.* **2001**, *114*, 1981.
- (92) Shamasundar, K. R.; Asokan, S.; Pal, S. *J. Chem. Phys.* **2004**, *120*, 6381.
- (93) Theis, D.; Khait, Y. G.; Pal, S.; Hoffmann, M. R. *Chem. Phys. Lett.* **2010**, *487*, 116.
- (94) Szalay, P. *Int. J. Quantum Chem.* **1995**, *55*, 151.
- (95) Ajitha, D.; Hirao, K. *Chem. Phys. Lett.* **2001**, *341*, 121.
- (96) Manohar, P. U.; Val, N.; Pal, S. *J. Mol. Struct. THEOCHEM* **2006**, *768*, 91.
- (97) Manohar, P. U.; Pal, S. *Chem. Phys. Lett.* **2007**, *438*, 321.
- (98) Bag, A.; Manohar, P. U.; Val, N.; Pal, S. *J. Chem. Phys.* **2009**, *131*, 024102.
- (99) Christiansen, O.; Jorgensen, P.; Hattig, C. *Int. J. Quantum Chem.* **1998**, *68*, 1.
- (100) Stanton, J. F. *J. Chem. Phys.* **1993**, *99*, 8840.
- (101) Stanton, J. F.; Gauss, J. *J. Chem. Phys.* **1994**, *101*, 8938.
- (102) Stanton, J. F.; Gauss, J. *J. Chem. Phys.* **1995**, *103*, 88931.
- (103) Nooijen, M.; Bartlett, R. J. *J. Chem. Phys.* **1997**, *107*, 6812.
- (104) Gwaltney, S. R.; Bartlett, R. J.; Nooijen, M. *J. Chem. Phys.* **1999**, *111*, 58.
- (105) Levchenko, S. V.; Wang, T.; Kyrlov, A. I. *J. Chem. Phys.* **2005**, *122*, 224106.
- (106) Scuseria, G. E. *J. Chem. Phys.* **1991**, *94*, 442.
- (107) Gauss, J.; Stanton, J. F. *Chem. Phys. Lett.* **1997**, *276*, 70.
- (108) Gauss, J.; Stanton, J. F. *J. Chem. Phys.* **2002**, *116*, 1773.
- (109) Harding, M. E.; Metzroth, T.; Gauss, J. *J. Chem. Theory Comput.* **2008**, *4*, 64.
- (110) Bak, K. L.; Gauss, J.; Helgaker, T.; Jorgensen, P.; Olsen, J. *Chem. Phys. Lett.* **2000**, *319*, 563.
- (111) Sattelmeyer, K. W.; Stanton, J. F.; Olsen, J.; Gauss, J. *Chem. Phys. Lett.* **2001**, *347*, 499.
- (112) Lindgren, I. *Int. J. Quantum Chem.* **1979**, *S12*, 33.
- (113) Kallay, M.; Gauss, J.; Szalay, P. *J. Chem. Phys.* **2003**, *119*, 2991.
- (114) Neogrady, P.; Medved, M.; Cernusak, I.; Urban, M. *Mol. Phys.* **2002**, *100*, 541.
- (115) Urban, M.; Neogrady, P.; Raab, J.; Diercksen, G. H. F. *Collect. Czech. Chem. Commun.* **1998**, *63*, 1409.
- (116) Thomson, R.; Dalby, F. F. *Can. J. Phys.* **1968**, *46*, 2815.
- (117) Huzinaga, S. *J. Chem. Phys.* **1965**, *42*, 1293.
- (118) Dunning, T. H. *J. Chem. Phys.* **1970**, *53*, 2823.
- (119) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007.
- (120) Sadlej, A. J. *Theor. Chim. Acta* **1991**, *79*, 123.
- (121) Phekp, D. H.; Dalby, F. W. *Phys. Rev. Lett.* **1966**, *16*, 3.
- (122) Kanata, H.; Yamamoto, S.; Saito, S. *J. Mol. Spectrosc.* **1988**, *131*, 89.
- (123) (a) Stanton, J. F.; Gauss, J.; Perera, S. A.; Watts, J. D.; Yau, A. D.; Nooijen, M.; Oliphant, N.; Szalay, P. G.; Lauderdale, W. J.; Gwaltney, S. R.; Beck, S.; Balkova, A.; Bernholdt, D. E.; Baeck, K. K.; Rozyczko, P.; Sekino, H.; Huber, C.; Pittner, J.; Cencek, W.; Taylor, D.; Bartlett, R. J. *ACES II*; University of Florida: Gainesville, FL. (b) Integral packages included are: Almof, J.; Taylor, P. R. *VMOL*; University of Florida: Gainesville, FL. (c) Taylor, P. *VPROPS* University of Florida: Gainesville, FL. (d) Helgaker, T.; H. J. Aa. Jensen, Jorgensen, P.; Olsen, J.; Taylor, P. R. *ABACUS* University of Florida: Gainesville, FL. (e) Schmidt, M. W.; Baldrige, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. J.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *HONDO/GAMESS* University of Florida: Gainesville, FL, 2006.

Linear Scaling Constrained Density Functional Theory in CONQUEST

Alex M. P. Sena,[†] Tsuyoshi Miyazaki,[‡] and David R. Bowler^{*,†}[†]London Centre for Nanotechnology and Department of Physics and Astronomy, University College London, Gower St, London, WC1E 6BT, United Kingdom[‡]National Institute for Materials Science (NIMS), 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan

ABSTRACT: The constrained density functional theory (cDFT) formalism is implemented in the linear scaling density functional theory (DFT) code CONQUEST. This will enable the simulation of electron-transfer processes in large biologically and technologically relevant systems. The Becke weight population scheme is chosen to define the constraint, as it enables force components to be calculated both analytically and efficiently in a linear scaling code. It is demonstrated that the imposition of a constraint is not affected by the truncation of the density matrix. Demonstration calculations are performed on charge-separated excited states in small biphenyl molecules, and cDFT is found to produce accurate energy and geometry changes for this system. The capability of the method is shown in calculations on poly phenylene-vinylene oligomers and a hydrated DNA 10-mer.

1. INTRODUCTION

Much research has been performed to elucidate how electron-transfer (ET) processes operate in biological situations, both to further understand their significance and to investigate the application of these biological roles to technological applications. Particularly well-studied examples are the processes of photosynthesis and respiration.^{1,2} The transfer of electrons between porphyrin cofactors in the electron-transfer chain of respiration has inspired attempts at creating porphyrin nanowires.^{3,4} While the photoinduced charge separation and recombination occurring in the photosynthetic reaction center have led to attempts to create artificial solar cells using porphyrin based systems.⁵

The difficulties in understanding the electronic structure of these large biological systems and technological analogues as well as the experimental challenges involved when investigating them have resulted in the need to study such ET processes in such systems computationally. Density functional theory (DFT)^{6,7} is a widely used method for computing the ground-state properties of molecules and solids. However it is unsuited to studying electron transfer in the biological systems described above. In these systems charge transfer is usually considered as a hopping event. For the hopping of an unpaired localized electron, $D^-A \rightarrow DA^-$, the presence of the self-interaction error^{8,9} can cause the electron to delocalize across multiple centers. This prevents the study of localized charge states. In addition, as DFT is a ground-state theory, the charge-separated configurations of a neutral system, $D^-A^+ \rightarrow D^+A^-$, which may be excited states of the system, are not usually accessible. Finally, studying any large biological system using DFT is challenging as these systems can contain many thousands of atoms. DFT's cubic scaling results in there being a limit of approximately 1,000 atoms to the size of systems that can be studied, even with massively parallel machines.

There exists a number of ways to at least partially remedy these problems. The delocalization of unpaired electrons can be reduced through the use of hybrid functionals,¹⁰ which reduce the self-interaction errors by introducing a portion of exact

exchange. There also exists fully self-interaction corrected implementations of DFT.¹¹ Excited states can be studied using developments, such as linear response time-dependent DFT.¹² For the system size problem, there are approaches which, using the locality of the density matrix, allow calculations to scale linearly with the number of atoms in the system.¹³ These codes can simulate systems with 10 000 atoms or more, although are often more difficult to apply in practice than conventional DFT codes. Incorporating the electron-transfer corrections described above into these linear scaling codes is challenging and may have a detrimental effect on the linear scaling. It would be desirable to have a linear scaling ET methodology to perform calculations on large biological systems.

Constrained density functional theory (cDFT)¹⁴ is another extension to DFT, able to rectify some of the difficulties faced when studying electron transfer. In this method, an extra potential is searched for that, when added to the Kohn–Sham Hamiltonian, imposes an experimentally or physically motivated constraint on the density. It was recently shown¹⁵ that cDFT can locate the energy minimum of the system with the addition of just one extra one-dimensional line minimization within each Kohn–Sham self-consistent step. This adds only moderate cost to a DFT calculation and does not affect the scaling with system size. A number of varied works have been performed using cDFT calculations.^{16–18}

In this paper, the implementation of the cDFT formalism into the linear scaling DFT code CONQUEST^{19–21} is described, and some initial calculations presented. First, the incorporation of cDFT within a linear scaling formalism is described. Then, test calculations are performed on small molecules which form the basis of future large-scale calculations to be performed. These include the hole localization in positively charged DNA base dimers and the structural changes occurring in charge separated biphenyl molecules.

Received: October 21, 2010

Published: March 07, 2011

2. METHODS

2.1. Linear Scaling DFT. Linear scaling DFT has been under development for around 15 years,²² and there are now a number of codes available, such as SIESTA,²³ ONETEP,²⁴ OPENMX,²⁵ and CONQUEST.²⁶ While there are many differences in technical implementations, the basic principles in each are similar. Conventional DFT minimizes an energy expression with respect to the density $n(\mathbf{r})$ and the Kohn–Sham orbitals $\psi_i(\mathbf{r})$. Linear scaling DFT codes recast the energy functional in terms of the density matrix, $\rho(\mathbf{r},\mathbf{r}')$ and the charge density $n(\mathbf{r}) = \rho(\mathbf{r},\mathbf{r}')$. Thus direct use of the Kohn–Sham orbitals is forgone. Instead of diagonalizing the Hamiltonian directly, the energy is minimized with respect to the density matrix for a given density. The process is repeated until the input and output densities are self-consistent (or using some other update scheme, mixing density matrix and charge updates). The density matrix is formally written as

$$\rho(\mathbf{r},\mathbf{r}') = \sum_i f_i \psi_i(\mathbf{r}) \psi_i(\mathbf{r}')^* \quad (1)$$

and the energy expression is written as

$$E[n,\rho] = E_{\text{KE}}[\rho(\mathbf{r},\mathbf{r}')] + E_{\text{PP}} + E_{\text{Har}} + E_{\text{xc}}[n(\mathbf{r}),\rho(\mathbf{r},\mathbf{r}')] \quad (2)$$

where the density is defined as $n(\mathbf{r}) = 2\rho(\mathbf{r},\mathbf{r})$. The terms in the energy are the usual DFT ones, respectively, the kinetic, the pseudopotential (representing electron–ion interaction), the Hartree, and the exchange–correlation energies. The Kohn–Sham orbitals can be expanded in terms of localized orbitals, often known as ‘support functions’ ϕ_α , so that $\psi_n = \sum_\alpha c_n^\alpha \phi_\alpha$ giving for the elements of the density matrix:^{27,28}

$$\rho(\mathbf{r},\mathbf{r}') = \sum_{\alpha\beta} \phi_\alpha(\mathbf{r}) K^{\alpha\beta} \phi_\beta(\mathbf{r}') \quad (3)$$

where $K^{\alpha\beta}$ can be formally defined as $K^{\alpha\beta} = \sum_n f_n c_n^\alpha c_n^\beta$. The near-sightedness of electronic matter²⁹ states that density correlations in a well gapped system falls off exponentially with distance. This is used to justify imposing a spatial cutoff on the density matrix such that $\rho(\mathbf{r},\mathbf{r}') \rightarrow 0$ as $|\mathbf{r} - \mathbf{r}'| \rightarrow \infty$. Consequently the number of nonzero elements in the density matrix and thus the computational cost can scale linearly with system size. The exact result is recovered as the density matrix cutoff and is increased.

Locality is imposed by restricting the support functions to lie within spherical regions and neglecting elements of the K matrix (either using a distance-based criterion, so that $K^{\alpha\beta} = 0$ when the centers of the support functions α and β are more than a specified distance apart or using a drop tolerance). The density matrix must be idempotent, to ensure that $K^{\alpha\beta}$ is a projector onto the occupied subspace during the minimization; this condition can be imposed in a variety of ways, but in the CONQUEST code, the LNV implementation of the McWeeny scheme^{30–32} is used. Here K is written in terms of an auxiliary density matrix, $L:K = 3LSL - 2LSLSL$, where S is the overlap matrix between support functions, $S_{\alpha\beta} = \langle \phi_\alpha | \phi_\beta \rangle$. The energy is minimized with respect to the elements of the auxiliary density matrix L using a standard scheme (either conjugate gradients or Pulay RMM-DIIS).

The calculations performed in this paper use a double- ζ + polarization pseudoatomic-orbital basis set (apart from the PPV and DNA in water simulations, which were tested with single-zeta and single-zeta plus polarisation basis sets)³³ and the Perdew, Burke and Ernzerhof (PBE) exchange correlation functional.^{34,35} Calculations can be performed using either exact

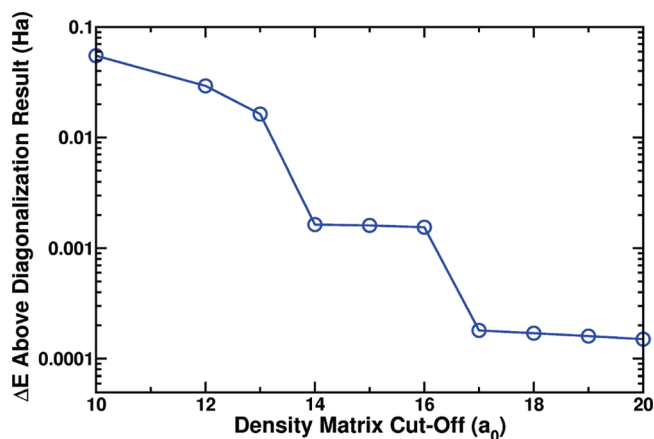


Figure 1. The convergence of energy from a $\mathcal{O}(N)$ calculation to that of a diagonalization calculation, as the density matrix cutoff is increased.

diagonalization or $\mathcal{O}(N)$ density matrix minimization. This is important as exact calculations on small systems can be performed to check accuracy of $\mathcal{O}(N)$ simulations.

2.2. Implementation of cDFT. The recent implementation of cDFT by Wu and Van Voorhis is now outlined. Full details are found elsewhere.^{15,36} It is desired to minimize the energy $E[n]$ of a system, subject to its density satisfying a constraint that the number of electrons in a given region of space around the system $w_c(\mathbf{r})$ is equal to a certain number N_c . Using the method of Lagrange multipliers, this is equivalent to minimizing a new functional W :

$$W[n, V_c] = E_{\text{KS}}[n] + V_c \left(\int w_c(\mathbf{r}) n(\mathbf{r}) d\mathbf{r} - N_c \right) \quad (4)$$

where the term in the bracket is the constraint and V_c the Lagrange multiplier. Minimizing W with respect to the density produces a Kohn–Sham orbital equation with an extra potential:

$$(H_{\text{KS}} + V_c w_c) \phi_i = \epsilon_i \phi_i \quad (5)$$

The minimum is located via self-consistent diagonalization of this new Hamiltonian. However at each step there is an inner loop in which the Lagrange multiplier is altered to impose the constraint on the density. In an $\mathcal{O}(N)$ scheme, only the expression for W is needed. It is recast to include the density matrix as

$$W[n, V_c] = E_{\text{KS}}[n] + V_c (2\text{Tr}[w_c K] - N_c) \quad (6)$$

The functional is minimized by self-consistent variation of the density matrix. However again at each step, there is an inner loop to alter the Lagrange multiplier and impose the constraint. In CONQUEST this is done using a Brent minimization algorithm to find the zero of $|\text{Tr}[K w_c] - N_c|$. The introduction of inner loops is relatively cheap, as the Hartree and exchange–correlation potentials do not need to be rebuilt each time. The inner loop also does not affect the linear scaling behavior of an $\mathcal{O}(N)$ code.

There are many different ways^{15,36} of defining the constraining potential $w_c(\mathbf{r})$. The representation of this spatial function in the basis of support functions turns the constraining potential $w_c(\mathbf{r})$ into a matrix $w_{\alpha\beta}^c$, known as the weight matrix. While there is no unambiguous way of apportioning a continuous electron density to atoms, or basis functions, there are a number of schemes which sensibly formulate the spatial potential $w_c(\mathbf{r})$ using this weight

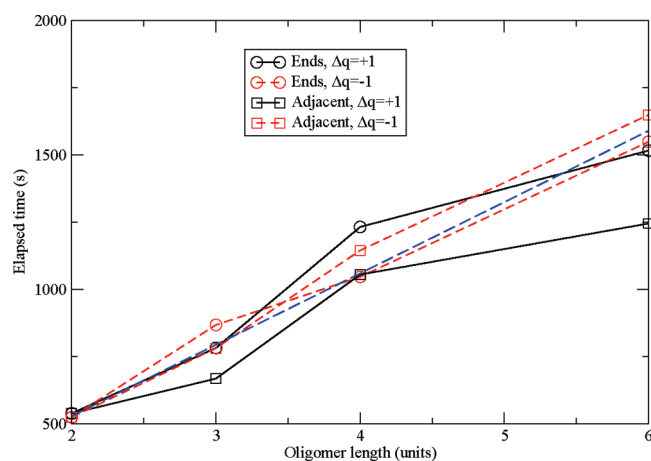


Figure 2. Total time required to find ground state for oligomers of PPV. Circles indicate charge confined to opposite ends of oligomer, and squares indicate charge confined to adjacent benzene rings. Tests were performed for differing polarities (solid and dashed lines). The long dashed line indicates linear scaling behavior extrapolated from the PPV dimer.

matrix. In our work a Becke weight scheme³⁷ is used:

$$w_{\alpha\beta}^{\text{Becke}} = \sum_{i \in C} \int d\mathbf{r} \phi_{\alpha}(\mathbf{r}) w_i^c(\mathbf{r}) \phi_{\beta}(\mathbf{r}) \quad (7)$$

This scheme has some advantages for use in a linear scaling code. It allows easy analytic calculation of force components, whereas schemes, such as the Löwdin populations, do not, as they require the gradient of the $S^{1/2}$ matrix. While there are schemes for finding these matrices,³⁸ we found them to be less practical than the Becke scheme; moreover, there can be problems decomposing S matrices with large basis sets.³⁸ A Cholesky decomposition does not parallelize well, and the CONQUEST code is massively parallel.

2.3. $\mathcal{O}(N)$ Calculation Convergence and Scaling. Here we present tests of the convergence of linear scaling cDFT to exact diagonalization and of the time taken to find the ground state when using linear scaling to solve.

To ensure that the truncation of the density matrix does not adversely affect the application of the constraint, a test calculation has been performed comparing the $\mathcal{O}(N)$ and diagonalization methods. cDFT is used to simulate a magnesium porphyrin molecule with the valence of the central atom constrained to be +1. Figure 1 shows how the calculated energy of the molecule for an $\mathcal{O}(N)$ cDFT calculation approaches that of a diagonalization cDFT calculation, as the range of the density matrix increases. Exact convergence occurs once the density matrix is larger than the maximum separation between two atoms in the system, approximately the diameter of the porphyrin ring here. The charge constraint was only solved to an accuracy of 10^{-4} , which will affect the total energy, as indicated in eq 6; therefore, differences between the exact diagonalization and the linear scaling solve of this order are expected, even at long density kernel ranges. We note that the accuracy is excellent for a density matrix cutoff greater than ~ 16 Å, which is in good agreement with recent calculations on DNA.³⁹

To test the effect on the scaling of the code when performing cDFT, a series of PPV (poly phenylene-vinylene) oligomers were constructed, and excitonic states were found using cDFT. Figure 2 shows the total time to the ground state for differing

Table 1. Binding Energies for DNA Base Dimers^a

dimer	separation (Å)	binding energy (eV)			
		DFT	cDFT	DFT (B3LYP)	SIC
adenine	2.50	-1.084	-2.019	-3.533	-3.450
	6.00	0.848	0.094	0.677	-0.182
guanine	2.50	-2.321	-3.412	-3.936	-4.418
	6.00	0.691	-0.082	0.538	0.030
cytosine	2.50	-0.308	-1.558	-2.348	-6.440
	6.00	1.001	0.015	0.812	0.017
thymine	2.50	-3.421	-4.564	-5.477	-6.302
	6.00	0.955	0.102	0.625	0.013

^aApplication of a constraining potential is found to bring the energies close to the SIC values.

exciton separations and polarities, with linear scaling behavior clearly shown. The ground-state search *without* cDFT for the dimer took 89 s and for the hexamer took 241 s, giving an overall increase in time by a factor of 6–7 for cDFT over simple DFT. This makes cDFT calculations on large systems accessible through linear scaling DFT.

3. RESULTS

An $\mathcal{O}(N)$ implementation of cDFT will enable electron transfer processes to be studied in large systems. However care must be taken when selecting a system for an $\mathcal{O}(N)$ calculation. Therefore as a first step, calculations are performed on small molecules which form the basis of interesting larger systems, one example of which is given at the end.

3.1. DNA Base Pairs. The natural self-assembly of DNA into a double helix makes it a potential candidate for use as a molecular wire.⁴⁰ However, there is currently no definitive consensus on the transport properties of DNA strands.⁴¹ Hole transfer along a DNA helix is known to be responsible for damage in the molecules, and there have been a number of studies showing high mobilities for charge carriers in DNA. However, other studies have also found DNA to be a poor charge transporter,^{42–44} while a recent study has indicated the importance of self-interaction.⁴⁵ A linear scaling implementation will allow investigation of large strands of DNA base pairs, surrounded by a solvent, and will answer some these questions; initial work³⁹ shows promise. Here some preliminary steps are made by investigating positively charged cofacial dimers of DNA bases. Of these, guanine dimers are particularly important as they possess a low ionization potential and thus often provide a pathway through which holes can travel in DNA.⁴⁶

When studying positively charged DNA dimers using DFT, the self-interaction error causes the hole wave function to be spread across both molecules.⁴⁷ A self-interaction corrected (SIC) DFT method has been used to localize the hole wave function on one base⁴⁷ and to investigate the change in binding energy that this produced. In this paper, the cDFT method is used to force the hole to localize on one of the bases. The binding energies for the constrained and unconstrained dimers are compared in Table 1 along with SIC values and those from a B3LYP calculation;⁴⁷ the inclusion of exchange can also mitigate self-interaction errors. The binding energies are generally negative, indicating unbound charges. We emphasize that, as cDFT is a different method to SIC, we expect only to see agreement of trends. In general, when the bases are well separated, cDFT shifts

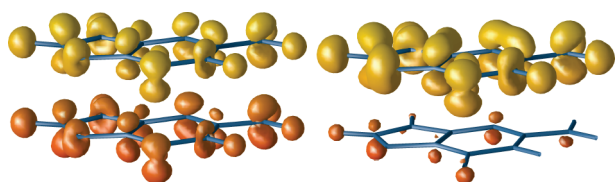


Figure 3. Charge difference density plot showing distribution of the positive hole in (left) DFT and (right) when constrained to be on the upper guanine. Color indicates height in the unit cell. The difference plotted is between the positively charged and the neutral systems.

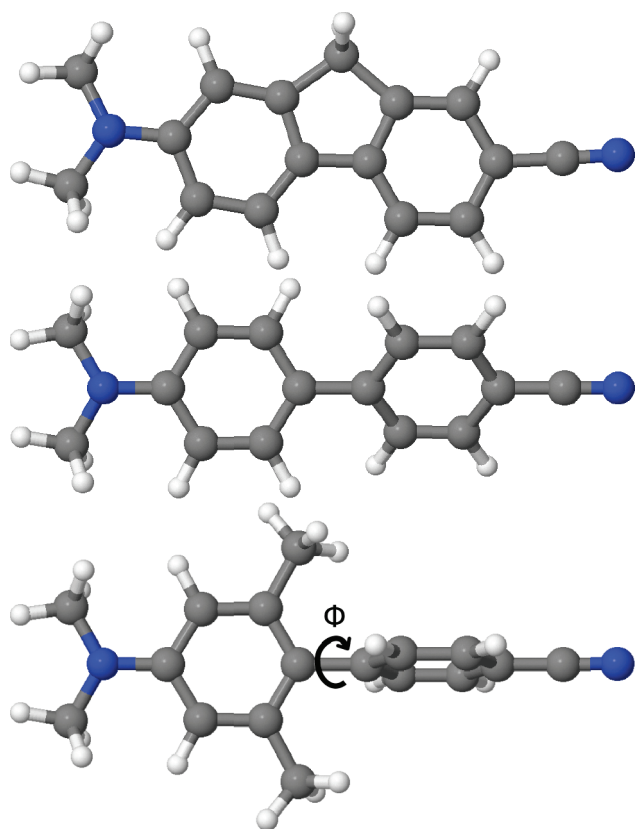


Figure 4. The structure of the three biphenyls used in the simulations here. The second (compound II) has a ground-state twist angle of 40° , and the third (compound III) has a ground-state twist angle of 70° .

the DFT binding energies toward the SIC values. However, when the bases are closer together, the cDFT results more closely resemble those from B3LYP calculations. This may be related to the problem of assigning charge to fragments which are close together,⁴⁸ though it is also quite possible that SIC methods are less accurate for close fragments.

Plotted in Figure 3 is the charge difference between a neutral and a positively charged guanine dimer for both DFT and cDFT cases. In the DFT case, the extra charge is spread across the whole dimer. When using cDFT, it is confined to just one base.

3.2. Shape Change of Biphenyls. There is considerable interest in finding molecules with switchable properties that could function as molecular electronic devices.^{49–51} One of the most sought after is a molecular photoswitch whose transport properties can be changed by the application of light.^{52,53} In this vein, an experimental study of three small biphenyl molecules has been made,⁵⁴ shown in Figure 4, with attention paid to the shape

Table 2. Change in Twist Angles When Charge Is Separated in Three Different Biphenyl Complexes

molecule	twist angle (ϕ)			
	AM1(Gr.St)	Expt(Ex.St)	DFT(G.St)	cDFT(Ex.St)
I	0°	0°	0°	0°
II	39°	0°	40°	0°
III	78°	40°	70°	40°

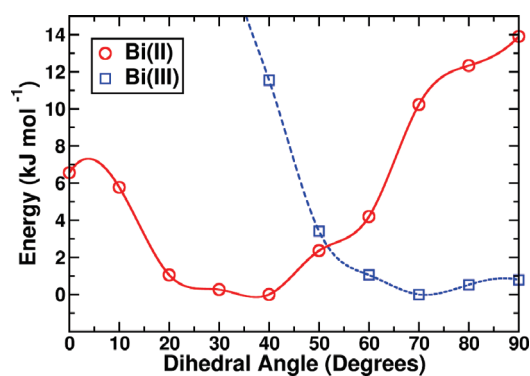


Figure 5. The twist potential (or energy profile of the twist angle) of the ground-state biphenyl II (blue line) and biphenyl III (red line) complexes.

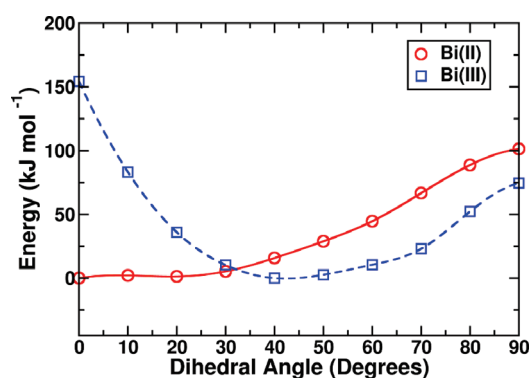


Figure 6. The excited-state twist potential for the biphenyl II and biphenyl III complexes. Complex II has an equilibrium twist angle of 0° , and III has an equilibrium angle of 40° .

changes that occur when the molecule is photoexcited from its ground state to a charge-separated state. Upon photoexcitation of the molecule, the benzonitrile group (RHS of molecule) was found to act as an acceptor and become negatively charged, while the dimethylaniline group (LHS of molecule) became the positive donor. The change this caused in the equilibrium dihedral (twist) angle ϕ was then investigated. The charge-separated twist potential was found to differ from the ground-state twist potential for molecules II and III, resulting in a different equilibrium angle as shown in Table 2. Molecule I is rotationally restricted. In molecule II, the excited state has a planar geometry as opposed the 39° twist angle of the ground state. In the gas phase, the excited state of compound III has a twist angle of 40° as opposed to its ground-state angle of around 78° in nonpolar solvents.

We have used CONQUEST to perform DFT and cDFT calculations on molecules II and III. Ground-state twist potentials

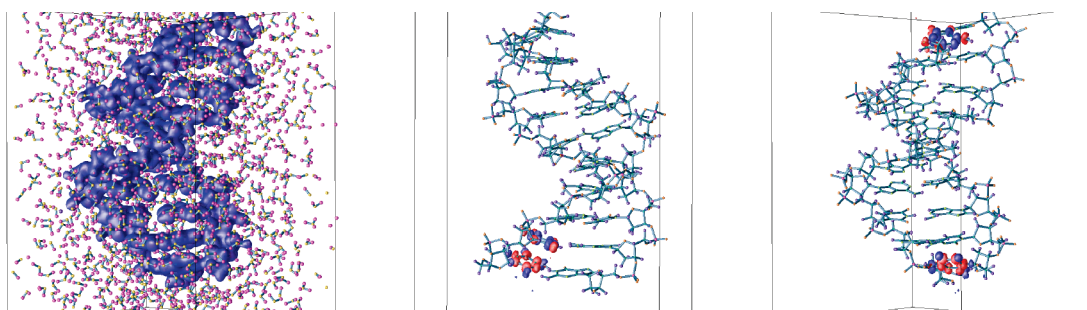


Figure 7. (a) The cell used showing water and DNA fragment with charge density of DNA fragment only shown for clarity. (b) Charge density difference (red for positive, blue for negative) for exciton confined to adjacent base pairs with DNA positions shown. (c) Charge density difference for exciton on separated base pairs with DNA positions shown.

were created by fixing the dihedral angle at various values and by allowing the other atoms of the molecule to relax. They are shown in Figure 5 and are similar to those produced using the AM1 method.⁵⁴ The twist potentials are extremely shallow, agreeing in size with previous work,⁵⁴ but the results are converged well enough to provide at least a qualitative indication of the ground-state angles, as summarized in Table 2 (column labeled DFT). cDFT was then used to create the charge separated (photoexcited) configurations. Again the twist potentials were calculated and are shown in Figure 6. This time the potentials are much deeper, and in both cases, the equilibrium angle is found to have decreased in a similar manner to the experimental observations. The new angles are seen in the column labeled cDFT in Table 2.

Having found the changes in angle, the reorganization energy for the transfer event can be calculated (though in our simulations the molecules were in the gas phase). The reorganization energy is defined as

$$\lambda = E(\mathbf{R}_{\text{Gr.St.}} n_D) - E(\mathbf{R}_{\text{CS.St.}} n_D) \quad (9)$$

for the separation event. By calculating the conventional DFT energy at the constrained geometry and by using the equation above, a value of $\lambda_{\text{cDFT}}^{\text{II}} = 6.50 \text{ kJ mol}^{-1}$ and $\lambda_{\text{cDFTRev}}^{\text{II}} = 15.74 \text{ kJ mol}^{-1}$ was found for molecule II, giving an average reorganization energy of $11.12 \text{ kJ mol}^{-1}$. For molecule III, $\lambda_{\text{cDFT}}^{\text{III}} = 11.55 \text{ kJ mol}^{-1}$ and $\lambda_{\text{cDFTRev}}^{\text{III}} = 22.99 \text{ kJ mol}^{-1}$, giving an average of $17.27 \text{ kJ mol}^{-1}$. The difference in the reorganization for charge separation and recombination is likely to be due to the fact that both the initial and final states are not exactly orthogonal diabatic states of the system and the potential energy surfaces not being perfectly harmonic. Values of the reorganization energy from previous work⁵⁴ can also be extracted using the experimental charge-separated twist potentials in conjunction with the theoretical AM1 results. These yield values of $\lambda_{\text{Expt}}^{\text{II}} = 8.6 \text{ kJ mol}^{-1}$ and $\lambda_{\text{Expt}}^{\text{III}} = 16.0 \text{ kJ mol}^{-1}$ in nonpolar solvents, which are most similar to the vacuum used in our simulations. The good agreement between these values and those found using cDFT is encouraging.

These results demonstrate that the energy and geometry of charge-separated excited states can be captured by creating the charge separation using cDFT. Although only a crude form of constraint altering the charge held on each half of the molecule has been used, the results are qualitatively a good fit to those from experiment. While the backward and forward reorganization energies vary significantly, their average is comparable with that from experiments on these molecules.

3.3. Charge Separation in DNA. We have previously characterized the performance of CONQUEST on a 10 base pair

strand of DNA in water,³⁹ with a total of 3439 atoms. Here we demonstrate cDFT calculations where we create and separate an exciton on the DNA strand.

Two calculations were performed, with the charges constrained to be on adjacent bases (separated by 3–4 Å) or at opposite ends of the strand (separated within the unit cell, along the strand, by 28–29 Å, and across the unit cell separated by water by 11–12 Å). The ground state was reached *without* cDFT in 7142 s running on 64 cores connected with Infiniband (an average of 53 atoms per core, distributed automatically using our Hilbert curve algorithm).⁵⁵ For the adjacent charges, the total time was 37 487 s (an increase of about a factor of 5), while for the larger separation the total time was 54 789 s (an increase of about a factor of 8), showing that even for this large system cDFT calculations are perfectly feasible; the variations in time reflect the different paths to the minimum energy for the two potentials. We show an illustration of the DNA strand and the charge density differences for the two excitons in Figure 7. The excellent convergence shown even for this large system demonstrates that the implementation is scalable and efficient. In future work, we will explore the energetics of charge separation in a variety of biological compounds.

4. CONCLUSIONS

The cDFT formalism has been implemented into the linear scaling DFT code CONQUEST. This will enable the study of electron-transfer events in large biologically and technologically relevant systems. The Becke weight population scheme was used to form the weight matrix, due to the ease with which both the weight matrix and the analytic force components can be calculated.

Demonstration calculations were performed on four different systems. First, linear scaling of constrained DFT was demonstrated for oligomers of PPV up to 90 atoms. Then, in positively charged dimers of DNA bases, cDFT was used to correct the delocalized nature of the hole distribution and was found to shift binding energies toward those found by a fully self-interaction-corrected DFT method.

The shape changes occurring upon creation of a charge-separated excited state in two biphenyl molecules were investigated. cDFT was found to reproduce accurately the experimentally observed changes in the dihedral angle caused by photoexcitation to a charge-separated state. Following these initial results, a demonstration of cDFT in a large system (3,439 atoms) was given for the creation and separation of an exciton in a hydrated DNA 10-mer.

Future studies will include constraining the charge on an ion as it passes through the gramicidinA ion channel and investigating

its affect on the ion channel structure. They will also focus on investigating charge separation in dye molecules deposited on TiO₂ surfaces,^{56,57} which have shown potential for use as artificial solar cells. It has recently been shown that linear scaling DFT calculations can address systems with millions of atoms,⁵⁸ which opens up the prospect of performing cDFT calculations on biologically relevant and important systems.

AUTHOR INFORMATION

Corresponding Author

*E-mail: david.bowler@ucl.ac.uk.

ACKNOWLEDGMENT

A.M.P.S. was supported by the IRC in Nanotechnology. D.R.B. was supported by the Royal Society. T.M. acknowledges the support from a Grant-in-Aid for Scientific Research from the MEXT and JSPS, Japan. The authors are grateful to Takao Otsuka for the positions of the DNA in water.

REFERENCES

- (1) Ermler, U.; Fritzsche, G.; Buchanan, S.; Michel, H. *Structure* **1994**, *2*, 925.
- (2) D'Souza, F.; Chitta, R.; Gadde, L., S.; Rogers, K.; Zandler, M.; Sandanayaka, A.; Araki, Y.; Ito, O. *Chem.—Eur. J.* **2007**, *13*, 916.
- (3) Tagami, K.; Tsukada, M.; Matsumoto, T.; Kawai, T. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2003**, *67*, 245324.
- (4) Kocherzhenko, A.; Patwardhan, S.; Grozema, F.; Anderson, H.; Siebbeles, L. *J. Am. Chem. Soc.* **2009**, *131*, 5522.
- (5) Campbell, W.; Burrell, A.; Officer, D.; Jolley, K. *Coord. Chem. Rev.* **2004**, *248*, 1363.
- (6) Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, B864.
- (7) Kohn, W.; Sham, L. *Phys. Rev.* **1965**, *140*, A1133.
- (8) Perdew, J. P.; Zunger, A. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1981**, *23*, 5048.
- (9) Zhang, Y.; Yang, W. *J. Chem. Phys.* **1998**, *109*, 2604.
- (10) Becke, A. J. *J. Chem. Phys.* **1993**, *98*, 5648.
- (11) Perdew, J. P.; Zunger, A. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1981**, *23*, 5048.
- (12) Marques, M.; Gross, E. *Annu. Rev. Phys. Chem.* **2004**, *55*, 427.
- (13) Goedecker, S. *Rev. Mod. Phys.* **1999**, *71*, 1085.
- (14) Dederichs, P. H.; Blügel, S.; Zeller, R.; Akai, H. *Phys. Rev. Lett.* **1984**, *53*, 2512.
- (15) Van Voorhis, T.; Wu, Q. *Phys. Rev. A: At., Mol., Opt. Phys.* **2005**, *72*, 024502.
- (16) Oberhofer, H.; Blumberger, J. *J. Chem. Phys.* **2009**, *131*, 064101.
- (17) Rudra, I.; Wu, Q.; Van Voorhis, T. *J. Chem. Phys.* **2006**, *124*, 024103.
- (18) Wu, Q.; Kaduk, B.; Van Voorhis, T. *J. Chem. Phys.* **2009**, *130*, 034109.
- (19) Bowler, D.; Miyazaki, T.; Gillan, M. *J. Phys.: Condens. Matter* **2002**, *14*, 2781.
- (20) Miyazaki, T.; Bowler, D.; Choudhury, R.; Gillan, M. *J. Chem. Phys.* **2004**, *121*, 6186.
- (21) Bowler, D. R.; Choudhury, R.; Gillan, M. J.; Miyazaki, T. *Phys. Status Solidi B* **2006**, *243*, 989.
- (22) Bowler, D. R.; Fattebert, J.-L.; Gillan, M. J.; Haynes, P. D.; Skylaris, C.-K. *J. Phys.: Condens. Matter* **2008**, *20*, 290301.
- (23) Artacho, E.; Anglada, E.; Dieguez, O.; Gale, J.; Garcia, A.; Junquera, J.; Martin, R.; Ordejon, P.; Pruneda, J.; Sanchez-Portal, D.; Soler, J. *J. Phys.: Condens. Matter* **2008**, *20*, 064208.
- (24) Skylaris, C.-K.; Haynes, P. D.; Mostofi, A. A.; Payne, M. C. *J. Chem. Phys.* **2005**, *122*, 084119.
- (25) Ozaki, T.; Terakura, K. *Phys. Rev. B* **2001**, *64*, 195126.
- (26) Gillan, M.; Bowler, D.; Torralba, A.; Miyazaki, T. *Comput. Phys. Commun.* **2007**, *177*, 14.
- (27) Hernández, E.; Gillan, M. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1995**, *51*, 10157.
- (28) Hernandez, E.; Gillan, M.; Goringe, C. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1995**, *53*, 7147.
- (29) Prodan, E.; Kohn, W. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *102*, 11635.
- (30) Li, X.-P.; Nunes, R.; Vanderbilt, D. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1993**, *47*, 16.
- (31) McWeeny, R. *Rev. Mod. Phys.* **1960**, *32*, 2.
- (32) Bowler, D. R.; Gillan, M. *J. Comp. Phys. Commun.* **1999**, *120*, 95.
- (33) Torralba, A.; Todorovic, M.; Brazdova, V.; Choudhury, R.; Miyazaki, T.; Gillan, M.; Bowler, D. *J. Phys.: Condens. Matter* **2008**, *20*, 294206.
- (34) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865.
- (35) Torralba, A.; Bowler, D.; Miyazaki, T.; Gillan, M. *J. Chem. Theory Comput.* **2009**, *9*, 1499.
- (36) Van Voorhis, T.; Wu, Q. *J. Chem. Theory Comput.* **2005**, *2*, 765.
- (37) Becke, A. J. *J. Chem. Phys.* **1988**, *88*, 2547.
- (38) Jansik, B.; Host, S.; Jorgenssen, P.; Olsen, J.; Helgaker, T. *J. Chem. Phys.* **2007**, *126*, 124104.
- (39) Otsuka, T.; Miyazaki, T.; Ohno, T.; Bowler, D.; Gillan, M. *J. Phys.: Condens. Matter* **2008**, *20*, 294201.
- (40) Priyadarshy, S. *Synth. React. Inorg. Met.* **2007**, *37*, 353.
- (41) Endres, R.; Cox, D.; Singh, R. *Rev. Mod. Phys.* **2004**, *76*, 195.
- (42) Guo, X.; Gorodetsky, A.; Hone, J.; Barton, J.; Nuckolls, C. *Nat. Nano.* **2008**, *3*, 163.
- (43) Delaney, S.; Barton, J. *J. Org. Chem.* **2003**, *68*, 6475.
- (44) Gomez-Navarro, C.; Moreno-Herrero, F.; de Pablo, P.; Colchero, J.; Gomez-Herrero, J.; Baro, A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *99*, 8484.
- (45) Pemmaraju, C. D.; Rungger, I.; Chen, X.; Rocha, A. R.; Sanvito, S. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2010**, *82*, 125426.
- (46) Kawai, K.; Kodera, H.; Osakada, T., Y.; Majima *Nat. Chem.* **2009**, *1*, 156.
- (47) Mantz, Y.; Gervasio, F.; Laino, T.; Parrinello, M. *J. Phys. Chem. A* **2006**, *111*, 105.
- (48) Wu, Q.; Cheng, C.-L.; Van Voorhis, T. *J. Chem. Phys.* **2007**, *127*, 164119.
- (49) Feldheim, D. *Nature* **2000**, *408*, 45.
- (50) Qian, H.; Lua, J.-Q. *Phys. Lett. A* **2007**, *371*, 465–468.
- (51) Ying Quek, S.; Kamenetska, M.; Steigerwald, M.; Joon Choi, H.; Louie, S.; Hybertsen, M.; Neaton, J.; Venkataraman, L. *Nat. Nano.* **2009**, *4*, 230–234.
- (52) Sallenaeva, X.; Delbaere, S.; Vermeersch, G.; Salehc, A.; Pozzo, J.-L. *Tetrahedron Lett.* **2005**, *46*, 3257–3259.
- (53) Sauer, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 9433.
- (54) Maus, M.; Rettig, W.; Bonafoux, D.; Lapouyade, R. *J. Phys. Chem. A* **1999**, *103*, 3388.
- (55) Brazdova, V.; Bowler, D. *J. Phys.: Condens. Matter* **2008**, *20*, 275223.
- (56) O'Rourke, C.; Bowler, D. R. *J. Phys. Chem. C* **2010**, *114*, 20240.
- (57) Terranova, U.; Bowler, D. R. *J. Phys. Chem. C* **2010**, *114*, 6491.
- (58) Bowler, D. R.; Miyazaki, T. *J. Phys.: Condens. Matter* **2010**, *22*, 074207.

On the Use of Accelerated Molecular Dynamics to Enhance Configurational Sampling in Ab Initio Simulations

Denis Bucher,* Levi C. T. Pierce, J. Andrew McCammon, and Phineus R. L. Markwick

Department of Chemistry and Biochemistry, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093-0365, United States

S Supporting Information

ABSTRACT: We have implemented the accelerated molecular dynamics approach (Hamelberg, D.; Mongan, J.; McCammon, J. A. *J. Chem. Phys.* **2004**, *120* (24), 11919) in the framework of ab initio MD (AIMD). Using three simple examples, we demonstrate that accelerated AIMD (A-AIMD) can be used to accelerate solvent relaxation in AIMD simulations and facilitate the detection of reaction coordinates: (i) We show, for one cyclohexane molecule in the gas phase, that the method can be used to accelerate the rate of the chair-to-chair interconversion by a factor of $\sim 1 \times 10^5$, while allowing for the reconstruction of the correct canonical distribution of low-energy states; (ii) We then show, for a water box of 64 H₂O molecules, that A-AIMD can also be used in the condensed phase to accelerate the sampling of water conformations, without affecting the structural properties of the solvent; and (iii) The method is then used to compute the potential of mean force (PMF) for the dissociation of Na–Cl in water, accelerating the convergence by a factor of ~ 3 –4 compared to conventional AIMD simulations.² These results suggest that A-AIMD is a useful addition to existing methods for enhanced conformational and phase-space sampling in solution. While the method does not make the use of collective variables superfluous, it also does not require the user to define a set of collective variables that can capture all the low-energy minima on the potential energy surface. This property may prove very useful when dealing with highly complex multidimensional systems that require a quantum mechanical treatment.

INTRODUCTION

In recent years, ab initio molecular dynamics (AIMD) has emerged as a promising tool for performing accurate free energy calculations from first principles.³ AIMD has been successfully applied to the study of a diverse variety of systems, including isolated molecules and condensed matter and solid-state systems.⁴ AIMD has also been employed in a quantum mechanics/molecular mechanics (QM/MM) manifold to investigate enzymatic reactions.⁵ However, the potential of AIMD to obtain accurate free energy statistics is hindered by the fact that configurational transitions or chemical reactions occur on time scales that are significantly longer than those accessible using standard AIMD methodologies. Despite the sustained and rapid increase in available computational power and the continued development of efficient simulation algorithms, AIMD simulations of even small, isolated molecules are generally limited to time scales of hundreds of picoseconds.

In the last two decades, considerable progress has been made in the development of more sophisticated methods to explore the configurational space of molecular systems more efficiently, allowing for the study of slow molecular motions and rare events. In general, these methods can be divided into two groups: The first involves the identification of transition pathways between known initial and final states. Such methods include, for example, transition path sampling,⁶ targeted molecular dynamics⁷ (and constrained dynamics⁸ in general), and essential molecular dynamics.⁹ The second group contains those methods that efficiently sample low-energy molecular conformations, allowing the rapid identification of thermodynamically dominant regions

on the potential energy surface (PES). These methods include replica exchange MD¹⁰ and metadynamics.¹¹

In the specific context of AIMD, the two most popular contemporary free energy methods employed are constrained MD and metadynamics. In the constrained MD method, a series of simulations are performed using a predefined internal degree of freedom as a constraint, and the free energy profile is obtained by integrating the average constraint force over the reaction coordinate. In metadynamics, the system is destabilized along a small set of predefined collective variables (internal degrees of freedom) by adding Gaussian potentials onto the PES in a history-dependent fashion. The free-energy surface is then obtained as the negative of the total bias potential added during the simulation. Noticeably, the successful application of both these enhanced sampling methods is dependent on the appropriate definition of a reaction coordinate or a set of collective variables and therefore requires at least some a priori understanding of the underlying PES.

In this paper we explore an alternative biased potential method that has been proposed recently in the framework of classical molecular dynamics, called accelerated Molecular Dynamics (aMD).¹ In the original variant of aMD, one adds a continuous non-negative bias potential to the actual PES, while still maintaining the essential details of the underlying PES. This has the effect of raising the low-energy regions on the potential energy landscape, decreasing the magnitude of energy barriers and

Received: October 24, 2010

Published: March 04, 2011

accelerating the exchange between low-energy conformational states. One of the favorable characteristics of this method is that one can recover the canonical average of an observable so that thermodynamic and other equilibrium properties can be accurately determined. In comparison to the other enhanced sampling methods described above, aMD does not require any prior knowledge of the underlying PES.

In the context of classical simulations, aMD has already been successfully employed to study slow time-scale dynamics in proteins, such as HIV-protease,¹² ubiquitin,¹³ IKBA,¹⁴ and H-Ras.¹⁵ The enhanced conformational space sampled by aMD has also been shown to significantly improve the theoretical prediction of experimental NMR observables, such as residual dipolar couplings, scalar J-couplings,¹³ and chemical shifts,¹⁴ which are sensitive to dynamic averaging on the micro- to millisecond time scale.

In this paper, we explore the possibility of using the accelerated ab initio MD (A-AIMD) to study conformational transitions and enhance phase-space sampling in the condensed phase. The present work is aimed at developing the A-AIMD method for studying systems in aqueous solution, which could initiate a variety of new applications, since most biological, chemical, and industrial processes, occur in water.

After presenting briefly the formal methodology, we present our results for three test systems: First, we investigate the conformational behavior of an isolated cyclohexane molecule in the gas phase. We demonstrate how A-AIMD can be used to explore the PES, identifying different molecular conformers, and affording accurate relative free energy statistics. Next, we present the results of A-AIMD simulations performed on bulk water, where it is shown that one can accelerate rotational and translational diffusion, and hence the sampling of water conformations, while still maintaining accurate free-energy weighted structural properties of the system. We then apply the method to Na–Cl in solution, to show that one can accelerate the convergence of the dissociation free energy profile of the two ions.

THEORY AND COMPUTATIONAL DETAILS

The details of accelerated molecular dynamics have been discussed previously in the literature.^{1,17} Following Voter's hyperdynamics scheme,¹⁸ a reference boost energy E_b is defined, which lies above the minimum of the PES. At each step in the simulation, if the potential energy $V(r)$ lies below this boost energy, a continuous non-negative bias potential $\Delta V(r)$ is added to the actual potential. The application of the bias potential raises the low-energy valleys and decreases the magnitude of energy barriers, while maintaining the essential details of the energy landscape. Explicitly, the modified potential $V^*(r)$ is defined as

$$V^*(r) = V(r) + \Delta V(r) \quad (1)$$

where the bias potential is defined as

$$\Delta V(r) = \frac{(E_b - V(r))^2}{E_b - V(r) + \alpha} \quad (2)$$

The extent of acceleration (i.e., how much the PES is raised and flattened) is determined by the choice of the boost energy (E_b) and the acceleration parameter (α). More aggressive acceleration can be achieved either by increasing E_b to flatten the potential or by decreasing the magnitude of α , which reduces the roughness of the potential. In practice, finding optimal

parameters require some testing. One usually chooses parameters so that the magnitude of fluctuations in ΔV during the simulations approximate the energy barriers. In many cases, the barrier heights are not known, and optimal parameters are found by holding one of the two parameters, while allowing the second parameter to evolve until the system starts exploring new regions of the phase space.

The forces acting on the nuclei are expressed as

$$F_{aMD} = - \frac{\partial(V(r) + \Delta V(r))}{\partial r} \\ = \frac{\partial V(r)}{\partial r} - \frac{\partial[(E_b - V(r))^2 / (E_b - V(r) + \alpha)]}{\partial r} \quad (3)$$

which can be reformulated as (for a derivation see the Supporting Information):

$$F_{aMD} = F_{MD} \cdot (1 + [(E_b - V(r))^2 / (E_b - V(r) + \alpha)]^2 \\ - 2(E_b - V(r) / (E_b - V(r) + \alpha))) \quad (4)$$

The bias potential as defined above ensures that the derivative of the modified potential will not be discontinuous at points where $V(r) = E_b$.

One of the favorable characteristics of this method is that it yields a canonical average of an observable, so that thermodynamic and other equilibrium properties can be accurately determined. The corrected canonical ensemble average of any given property, $\langle A \rangle_c$ is obtained by reweighting each point in the configuration space on the modified potential by the strength of the Boltzmann factor of the bias energy, $\exp(\beta \Delta V(r, t_i))$, at that particular point:

$$\langle A \rangle_c = \frac{\int A \exp(-\beta V^*(r)) \exp(-\beta \Delta V(r)) dr}{\int \exp(-\beta V^*(r)) \exp(-\beta \Delta V(r)) dr} \\ = \frac{\int A \exp(-\beta V(r)) dr}{\int \exp(-\beta V(r)) dr} \quad (5)$$

Computational Details. AIMD simulations were performed using the Car–Parrinello (CP) scheme,¹⁹ and an in-house modified version of the CPMD 3.13 code.²⁰ The three systems studied in this work were: (i) an isolated cyclohexane molecule in the gas phase placed at the center of a cubic box of length $L = 12.00$ Å; (ii) a periodically repeating cubic box of length $L = 12.44$ Å containing 64 H₂O molecules; and (iii) a periodically repeating cubic box of length $L = 12.35$ Å containing 62 H₂O molecules, one Na⁺ ion and one Cl⁻ ion. The electronic structure problem was solved with density functional theory (DFT), and in each case, the Becke (B) exchange and Lee–Yang–Parr (LYP) correlation functional were employed.²¹ Although a variety of empirical corrections have been suggested to include the effect of dispersion forces in the BLYP functional, we chose to use the standard functional, which allows us to compare our results directly to previous works on these systems. For each system, a fictitious electron mass of 400 au was ascribed to the electronic degrees of freedom, and the coupled equations of motion were solved using the velocity Verlet algorithm²² with a time-step of 4 au core electrons and were treated using the norm-conserving pseudopotentials of Troullier and Martins,²³ and the valence orbitals were expanded in a plane-wave basis set up to an energy cutoff of 80 Ry. All standard and A-AIMD

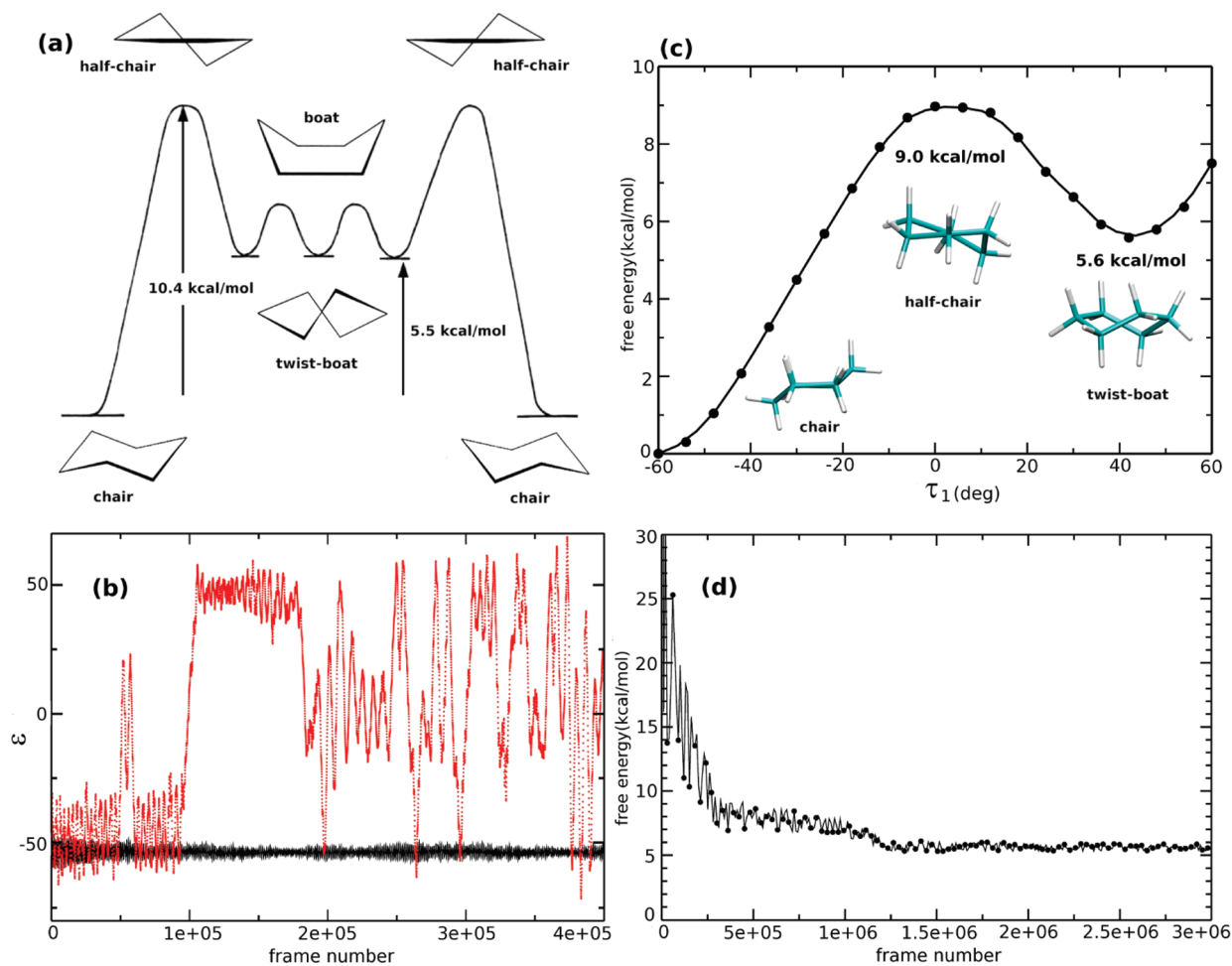


Figure 1. (a) Scheme representing the different conformations of cyclohexane. (b) The chair-to-chair interconversion was monitored during the simulations using a ε coordinate defined in the text. The A-AIMD simulation is shown in red (dotted line), and conventional AIMD in black (solid line), ($4e + 05$ frames = ~ 40 ps). (c) Constrained MD was used to calculate a reference PMF and to obtain the theoretical relative free energy of the twist–boat conformation with BLYP. (d) The A-AIMD estimate for the relative free energy of the twist–boat conformation converges toward the expected value.

simulations were performed at $T = 300$ K using a Nose–Hoover chain thermostat²⁴ on the ions with coupling frequency of 600 cm^{-1} . In the case of cyclohexane, a thermostat was also used on the electrons, with a target kinetic energy of 0.001 au and a coupling frequency of 6000 cm^{-1} .

RESULTS

Conformations of Cyclohexane. Accelerated AIMD is a highly efficient and robust conformational space sampling method. In order to demonstrate this, we performed an initial study to explore the PES of cyclohexane. Cyclohexane can exist in a variety of conformational states that have been depicted diagrammatically in Figure 1a. The most stable chemical conformation of cyclohexane is the ‘chair’ form, for which there are two geometric isomers. These isomers differ in respect to which hydrogen atoms in the ring adopt axial and equatorial positions. During the interconversion process, known as ‘ring flipping’, the axial hydrogens become equatorial and the equatorial hydrogens become axial, and the system passes through a metastable state, referred to as the ‘twist–boat’ conformation, which also possesses several geometric isomers. The transition states associated with interconversion between twist–boat isomers and the twist–boat and chair conformations are called

the ‘boat’ and ‘half-chair’ forms, respectively. Experimental studies have shown that the free energy of the twist–boat conformer lies approximately 5.5 kcal/mol above that of the thermodynamically stable chair conformation in the gas phase,²⁵ and the free energy barrier for interconversion between the chair and twist–boat conformers has been estimated by NMR experiments to be approximately 10.4 kcal/mol .²⁶ The ring-flipping process is therefore very slow, occurring on the microsecond time scale at 300 K , which is inaccessible using standard AIMD methods.

An initial standard AIMD simulation of cyclohexane was performed at $T = 300\text{ K}$ starting in the chair conformation. Unsurprisingly, this initial $400\,000$ steps (~ 40 ps) simulation confirmed that the chair conformation is very stable, and the average potential energy was -41.14 au ($1\text{ au} = 627.51\text{ kcal/mol}$). This value was used as the reference potential energy V_0 in the subsequent A-AIMD simulations. Keeping the acceleration parameter α fixed at 0.016 au ($\sim 10\text{ kcal/mol}$), a series of short ($400\,000$ steps) accelerated simulations were performed using different boost energies: $[E_b - V_0] = (0.02, 0.04, 0.06, 0.08, \text{ and } 0.10)$. As the value of $[E_b - V_0]$ was systematically increased, enhanced conformational space sampling was observed. Inspection of the resulting trajectories revealed that the ‘optimal’ acceleration parameters for observing the ring

flipping process within the time scale of the simulations were: $\{[E_b - V_0], \alpha\} = \{0.10 \text{ au}, 0.016 \text{ au}\}$.

The conformational changes occurring in the A-AIMD simulations were analyzed using geometric criteria defined by the dihedral angles of the ring (Table S11, Supporting Information). In addition, a configuration coordinate, ε , was formulated as

$$\varepsilon = (\tau_1 - \tau_2 + \tau_3 - \tau_4 + \tau_5 - \tau_6)/6 \quad (6)$$

where τ_i is the internal ring dihedral angle (see Supporting Information for more details). Using this configuration coordinate, the two isomeric forms of the chair conformation correspond to ε values of approximately -60 and $+60^\circ$ and the twist-boat intermediate corresponds to a value of ε of approximately 0. In Figure 1b, we show the conformational space sampling afforded by the extended 400 000 step 'optimal' A-AIMD simulation, compared to a standard AIMD simulation performed under the same physical conditions. In comparison to the standard AIMD simulation, which remains in the initial chair conformation throughout the entire trajectory, the optimal A-AIMD simulation readily interchanges between the different conformational states. Indeed, in the course of the 400 000 step (~ 40 ps) simulation, eight ring-flipping events were observed, and the system visited all known stable, metastable, and transition states (chair, half-chair, twist-boat, envelope, and boat, see Table S11, Supporting Information). The efficiency of the sampling is demonstrated by the fact that even within just 100 000 steps of A-AIMD (the equivalent of 10 ps), we observe a conformational transition which according to transition-state theory (assuming a transmission coefficient of 1) would actually occur on a time scale of approximately $1 \mu\text{s}$.

As discussed in the Theory and Computational Details Section, aMD is not only an efficient conformational space sampling algorithm but is also a robust free energy sampling method. However, obtaining accurate free energy statistics is more challenging than just exploring the conformational space: Relative free energy statistics are determined not only by the variation in the magnitude of the bias-potential but also by the density of states (i.e., the effective population) on the *modified potential*. Accurate free energy statistics can only be obtained if multiple transitions between the different conformational states are observed. In light of this, a second, much longer, 3 000 000 step (~ 300 ps) accelerated AIMD simulation was then performed at the optimal acceleration level. The relative free energy of the twist-boat conformer with respect to the chair conformation was calculated as $\Delta G_{(\text{twist-boat})} = -kT \ln(P_{(\text{twist})}/P_{(\text{chair})})$, using the density of the two states after reweighting the trajectories by the strength of the Boltzmann factor of the bias $[\exp(\beta\Delta V(r, t_i))]$. In order to obtain an accurate estimate of the *theoretical* free energy barrier for the chair to twist-boat conformational transition, using the BLYP density functional, we employed the well-established constrained MD approach (see Supporting Information for more details). The results of this study are presented in Figure 1c. The theoretical free energy barrier was found to be 9.0 ± 0.3 kcal/mol, which is in good agreement with the experimental estimate (10.4 kcal/mol). The A-AIMD simulation determined the free energy for the twist-boat conformation to lie 5.8 kcal/mol above that of the chair conformer, which is in excellent agreement with the reference constrained MD result (5.6 kcal/mol ± 0.3 kcal/mol, Figure 1c). The number of A-AIMD steps required to reach convergence in the free energy statistics was approximately 1.5×10^6 (see Figure 1d). In terms

of required CPU time, this is the equivalent of ~ 150 ps of standard AIMD, which is readily accessible.

In addition to obtaining the relative free energies of stable and metastable states, the magnitude of the free energy barrier can also be estimated approximately from A-AIMD simulations by measuring the amount of destabilization required to observe the conformational transition (i.e., the maximum bias potential, ΔV_{max}). The energy barrier for the chair-to-chair interconversion was estimated in this way by gradually increasing E_b (with α fixed) until the conformational transition was observed. A bias potential of ~ 10 kcal/mol was required to leave the chair conformation, in good agreement with the reference calculation for the energy barrier and with experimental data. Taken together, these encouraging results suggest that A-AIMD represents a useful method to sample the conformational space of isolated molecules. Next, we show that A-AIMD can also be used to accelerate the sampling for systems in the condensed phase. We have chosen for this study to carry out simulations on a water box. This choice was made not only because water is important in chemistry and biology but also because water is an important system from a theoretical perspective, as it is often used to test AIMD schemes and assist the development of new DFT functionals.

Enhanced Sampling in Condensed Matter Systems: Bulk Water. The cyclohexane study described above readily demonstrates how A-AIMD simulations can be used to efficiently explore the conformational space of isolated molecules and obtain accurate free energy statistics. However, such is the versatility of this method, A-AIMD simulations can also be employed to enhance the phase-space sampling in condensed matter systems. In order to investigate this, we performed a test study on bulk water. The particular focus of this study was to identify if it is possible to enhance the phase-space sampling using A-AIMD while still maintaining an accurate (free energy weighted) representation of the structural properties of the system.

Analogous to the cyclohexane study, an initial standard AIMD simulation was performed on a cubic box containing 64 water molecules under periodic boundary conditions, for 20 ps. The average density functional energy for the system was -1098.22 au, which was used as the reference potential energy, V_0 . A series of five 150 000 steps A-AIMD simulations were then performed at different acceleration levels. The specific acceleration parameters, $[E_b - V_0]$, and α , are presented in Table 1. As discussed in the Theory and Computational Details Section, the level of acceleration is determined by the relative magnitude of both $[E_b - V_0]$ and α . The A-AIMD simulations performed here, which we refer to as sim1–sim5, are ranked according to the level of acceleration determined by the average magnitude of the effective bias potential (ΔV_{ave} in Table 1).

The most direct way to assess the amount of phase-space sampling in condensed matter systems is to monitor the average translational and rotational diffusion properties of the composite molecules. In Figure 2a we show the mean square displacement of water molecules (computed after correcting for the displacement of the center of mass of the box) for the standard AIMD simulation and the five accelerated A-AIMD simulations. The effect of the bias potential clearly enhances the average translational diffusion of the water molecules. Compared to the standard AIMD simulation, the mean square displacement was found to increase from two- to eight-fold from the least aggressive (sim1) to most aggressive (sim5) A-AIMD simulation. In a

Table 1. Summary of the A-AIMD Simulations Performed for a Water Box with 64 H₂O Molecules^a

simulations	$E_b - V_0$ (au)	α (au)	time (ps)	ΔV_{ave} (kcal/mol)	ΔV_{max} (kcal/mol)	D_{acc}/D	τ/τ_{acc}
MD	—	—	20	—	—	1.0	1.0
sim1	0.1	0.1	15	0.8	5.4	2.1	1.7
sim2	0.2	0.4	15	1.1	7.3	2.9	1.9
sim3	0.2	0.1	15	1.7	11.2	4.3	2.1
sim4	0.3	0.4	15	2.5	15.4	6.4	3.0
sim5	0.3	0.1	15	3.1	16.6	7.1	3.6

^aThe two parameters E_b and α are used to control the level of acceleration. The difference between E_b and the average potential energy V_0 of a conventional AIMD simulation is given. The average and maximum values of the effective bias potential during the simulation are shown together with the acceleration with respect to conventional simulations in the observed diffusive properties (D_{acc}/D) and orientational correlation times (τ/τ_{acc}).

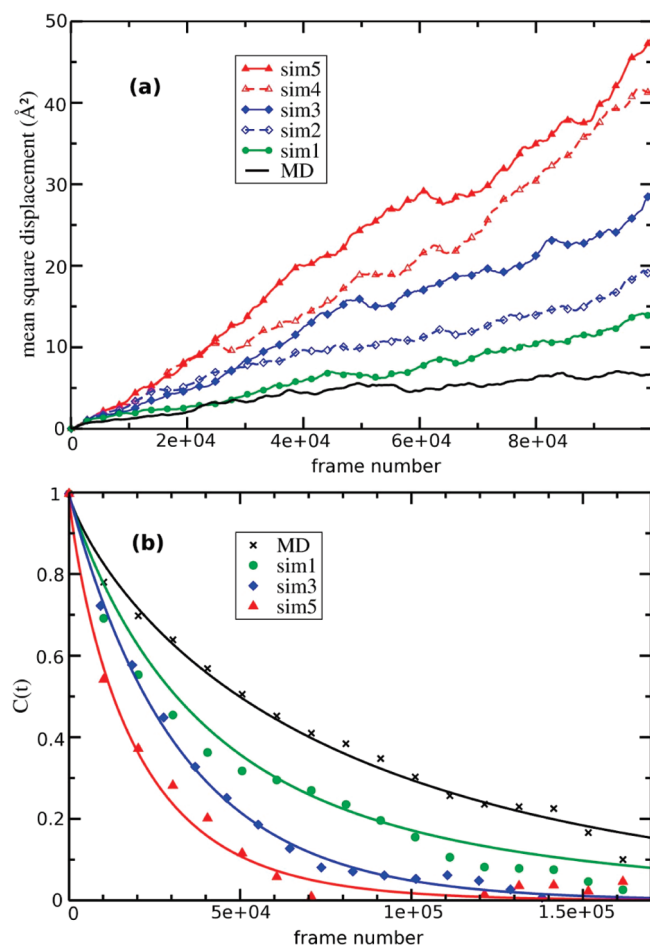


Figure 2. Observed dynamical properties of waters in conventional AIMD and A-AIMD simulations: (a) mean-square displacement and (b) orientational autocorrelation functions for O–H vectors.

similar manner, the application of the bias potential also increased the observed average rotational diffusion, which was assessed by calculating the effective (unweighted) reorientational autocorrelation function of the normalized O–H bond vector, averaged over all water molecules in the system. The reorientational autocorrelation functions shown in Figure 2b describe how a water molecule on average loses memory of its orientational ‘state’ during the simulation. As reorientation diffusion in condensed matter systems is a stochastic process, the associated autocorrelation functions show an exponential decay. By definition, in the limit that the autocorrelation function approaches

zero, the water molecules in the system have undergone a rotation of 2π radians on average. As can be seen in Figure 2b, even under only moderate acceleration (sim3), the reorientational correlation function approaches zero within 150 000 steps, compared to the standard AIMD simulation, where the reorientational correlation function approaches zero at step 315 000 (by extrapolation).

We would like to point out that during an A-AIMD simulation, the system evolves on a nonlinear time scale. Unfortunately, obtaining an accurate estimate of the time scale of the observed phase-space sampling in the A-AIMD simulations is not trivial. In light of this, we did not attempt to extract a meaningful estimate of the true translational and reorientational diffusion coefficients from the biased potential AIMD simulations. However, by comparing the effective mean-square displacements and reorientational relaxation ‘times’ (as a function of the number of MD steps), we can assess the effective enhancement in the translational and reorientational phase-space sampling compared to the standard AIMD simulation, reported in Table 1.

The results presented in Figure 2 clearly show that the effect of the bias potential significantly enhances the translational and reorientational diffusion properties of the system, and therefore A-AIMD simulations afford a substantial increase in the observed phase-space sampling. In order to study the structural properties of the system under the application of the bias potential, we calculated the free-energy weighted radial (O···O) and angular (H–O···O) distribution functions for all five A-AIMD simulations.

The O···O radial distribution functions are presented in Figure 3a and b along with the results obtained for both the standard AIMD simulation (20 ps) and an experimental X-ray diffraction study. Up to a moderate acceleration level (sim3), the free energy weighted radial distribution functions are in excellent agreement with the experimental X-ray diffraction data. The radial distribution functions obtained from the more aggressive accelerated simulations (sim4 and particularly sim5) appear to be less accurate, which is a direct result of the relatively short length of the simulations and the effect of enhanced statistical noise in the free energy reweighting protocol. When performing longer A-AIMD simulations at elevated acceleration levels, the noise in the free energy statistics will start to cancel out, and there may be in fact no loss in accuracy. The free energy weighted H–O···O angular distribution functions for the five A-AIMD simulations discussed here are presented in Figure 3c and compared to the standard AIMD result and experimental NMR data. Interestingly, a closer resemblance was obtained between the computed angular distribution function and the experimental NMR result as the level of acceleration (and therefore the extent of

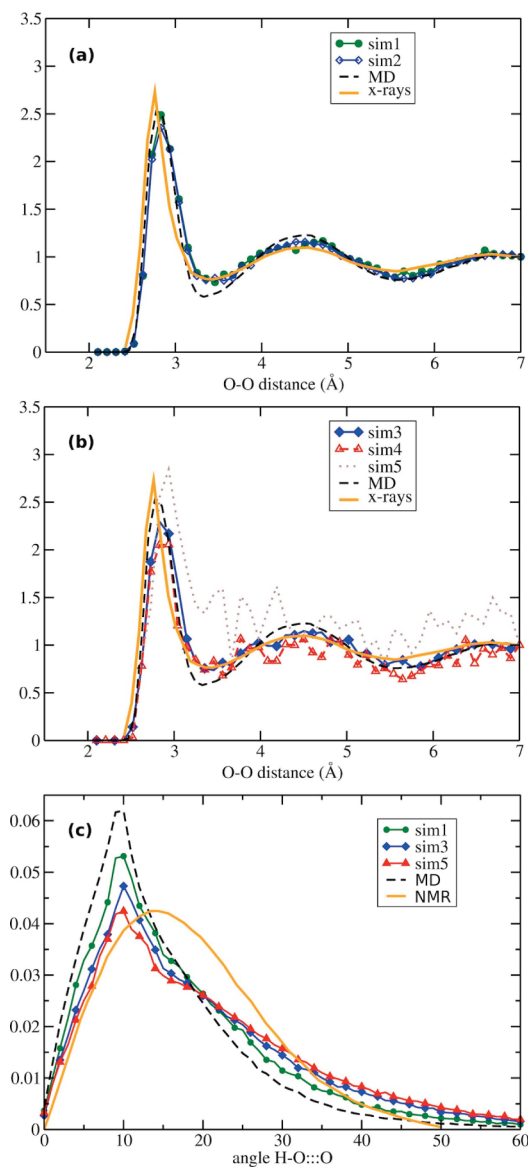


Figure 3. Structural properties of water in conventional AIMD and A-AIMD simulations. (a and b) Showing a comparison with X-rays diffraction²⁷ for the O–O radial distribution functions (RDFs) and (c) a comparison with NMR²⁸ for the orientation of hydrogen bonds.

phase-space sampled) was increased. We note in passing that under more aggressive acceleration conditions, we observed proton transfer events caused by the dissociation of water (data not shown). Although we did not attempt to study water autoionization here, it suggests that an even larger boost may be useful in some cases to study these rare events. Experimentally, a single water molecule is known to undergo autoionization in ~ 10 h.²⁹

Potential of Mean Force (PMF) Calculation for the Dissociation of Na–Cl in Water. The results obtained from our initial study on bulk water show that the application of the bias potential allows for a significant enhancement in the phase-space sampling, while still maintaining an accurate free energy weighted representation of the structural properties of the system up to moderate acceleration levels. In light of these results, it appears that A-AIMD can be employed as an efficient method to determine thermodynamic and equilibrium properties in

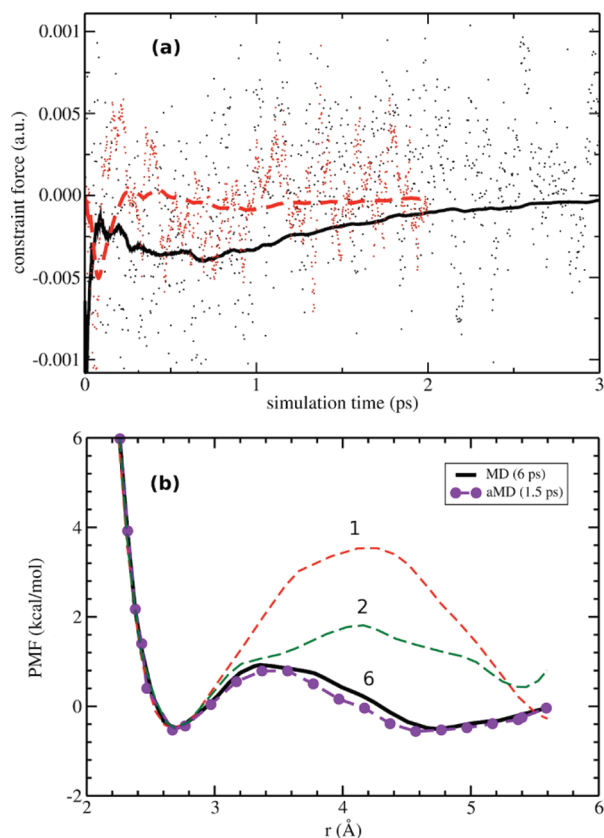


Figure 4. Convergence of the dissociation profile of NaCl in solution: (a) Cumulative average of the mean force at a separation length of 3.5 Å between Na and Cl (after reweighting the trajectories with eq 5). The convergence is shown for A-AIMD (dashed red line) and for conventional AIMD (solid black line). (b) Potential of mean force for the dissociation of NaCl in water computed with A-AIMD (1.5 ps per point) and with conventional AIMD (1, 2, and 6 ps per point).

condensed matter systems, particularly when these properties are sensitive to the effects of time and ensemble averaging, mediated by the diffusive properties of the solvent. In order to demonstrate this, we have calculated the free energy profile for the dissociation of Na–Cl in bulk water using an extended accelerated AIMD approach. Classical MD studies of this system date to 1984.³⁰ However, it has been studied only recently with ab initio methods.²

The simulation details and setup employed in this work are identical to those of the recently published study by Timko et al. using conventional constrained AIMD simulations. In this published study, the reaction coordinate, r , was defined as the distance between the Na and Cl ions, and it was found that a simulation time of up to 6 ps was required in order to converge the average constraint force at each constraint distance. The free energy profile (or potential of mean force) for Na–Cl dissociation was then obtained by integrating the average constraint force over the reaction coordinate. In the present work, we have performed the same constrained simulations in the framework of A-AIMD, using the ‘optimal’ acceleration parameters obtained from the bulk water study presented above: $\{E_b, -V_0, \alpha\} = \{0.2 \text{ au}, 0.1 \text{ au}\}$.

In Figure 4a we compare the convergence of the cumulative average constraint force obtained from A-AIMD to that obtained using standard AIMD for a constraint distance, $r = 3.5$ Å. The

cumulative average of the mean force in the A-AIMD simulation (which is free energy weighted) converges to the same value as in the conventional AIMD simulation (-0.0004 au) within 15 000 steps (the equivalent of 1.5 ps of standard AIMD simulation). Using this result as a guideline, the free energy profile for dissociation of NaCl in bulk water was calculated by performing 21 constrained accelerated AIMD simulations for 15 000 steps across the entire reaction coordinate from $r = 2.7$ to 5.7 Å (Figure 4b). The resulting potential of mean force is in excellent agreement with the previously published results of Timko et al., who used a sampling time of 6 ps. Therefore, the application of the bias potential affords an approximate four-fold speed up in the convergence of the potential of mean force, as is readily demonstrated in Figure 4b. This four-fold speed-up is consistent with the estimated enhanced phase-space sampling results obtained from the comparative analysis of the translational and reorientational diffusion properties of bulk water (Table 1, sim3).

CONCLUSIONS

We have discussed the preliminary testing and implementation of the accelerated MD approach in the framework of AIMD. Using three simple examples, we have demonstrated that A-AIMD is a highly efficient and robust method for enhanced conformational and phase-space sampling. In particular, we have shown that the effect of the bias potential allows for the study of slow conformational transitions and rare events, such as the ring-flipping process in cyclohexane, that occurs on microsecond time scales and is inaccessible when using standard AIMD. For isolated molecules, we have shown that A-AIMD affords accurate free energy statistics that allows for the determination of thermodynamic and other equilibrium properties. In the condensed phase, obtaining converged free energy with A-AIMD may be more challenging, due to the requirement that rare events are sampled many times. However, convergence problems can be avoided by using a different free energy method that allows for the computation of the free energy along collective variables. As an example, we have shown that A-AIMD can be used in conjunction with constrained MD to accelerate the convergence of constrained MD by a factor of ~ 4 , for the case of 2 ions in a water box. The A-AIMD method is likely to be a useful addition to existing methods for sampling purposes and for helping to determine an optimal set of collective variables that can describe all the low-energy transformations.

As mentioned in the Introduction, one interesting feature of A-AIMD is that it can accelerate rare events while maintaining the essential details of the underlying PES. Since the ordering of minima on the PES is conserved, events that are low in energy are likely to occur first during an A-AIMD simulation. This property can be used to gain an intuitive understanding of chemical reactivity. In future applications, we plan to apply the method to study complex chemical reactions in the condensed phase.

Finally, although all the biased potential simulations presented here have been performed in the framework of Car–Parrinello MD, we would like to note that the implementation of this method in the framework of Born–Oppenheimer or Ehrenfest dynamics is equally viable. In conclusion, A-AIMD represents a highly efficient and versatile addition to existing ab initio methodologies for performing enhanced conformational space sampling and determining accurate free energies. In future works, A-AIMD could also be used to study chemical reactions in solution or applied to larger systems of biological relevance within a QM/MM framework.

ASSOCIATED CONTENT

S Supporting Information. Derivation of the formula for the forces is given, the geometrical criteria used to analyze the cyclohexane trajectories are presented, together with computational details about the constrained MD calculations. This information is available free of charge via the Internet at <http://pubs.acs.org/>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: bucher.denis@gmail.com.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation, the National Institutes of Health, the Howard Hughes Medical Institute, the Center for Theoretical Biological Physics, the National Biomedical Computation Resource, the National Science Foundation Supercomputer Centers, and the Swiss Science Foundation (DB).

REFERENCES

- (1) Hamelberg, D.; Mongan, J.; McCammon, J. A. Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *J. Chem. Phys.* **2004**, *120* (24), 11919.
- (2) Timko, J.; Bucher, D.; Kuyucak, S. Dissociation of NaCl in water from ab initio molecular dynamics simulations. *J. Chem. Phys.* **2010**, *132* (11), 114510.
- (3) Leone, V.; Marinelli, F.; Carloni, P.; Parrinello, M. Targeting biomolecular flexibility with metadynamics. *Curr. Opin. Struct. Biol.* **2010**, *20* (2), 148.
- (4) Hutter, J.; Marx, D. *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods*; Cambridge University Press: Cambridge, U.K., 2009; p 578.
- (5) Senn, H. M.; Thiel, W. QM/MM Studies of Enzymes. *Curr. Opin. Chem. Biol.* **2007**, *11* (2), 182. (b) Laio, A.; VandeVondele, J.; Rothlisberger, U. A Hamiltonian electrostatic coupling scheme for hybrid Car–Parrinello molecular dynamics simulations. *J. Chem. Phys.* **2002**, *116* (16), 6941.
- (6) Dellago, C.; Bolhuis, P. G.; Csajka, F. S.; Chandler, D. Transition path sampling and the calculation of rate constants. *J. Chem. Phys.* **1998**, *108* (5), 1964.
- (7) Schlitter, J.; Engels, M.; Kruger, P.; Jacoby, E.; Wollmer, A. Targeted Molecular-Dynamics Simulation of Conformational Change Application to the T-R Transition in Insulin. *Mol. Simul.* **1993**, *10* (2–6), 291.
- (8) Ciccotti, G.; Ferrario, M.; Hynes, J. T.; Kapral, R. Constrained Molecular-Dynamics and the Mean Potential for an Ion-Pair in a Polar-Solvent. *Chem. Phys.* **1989**, *129* (2), 241.
- (9) Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. Essential Dynamics of Proteins. *Proteins: Struct., Funct., Genet.* **1993**, *17* (4), 412.
- (10) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314* (1–2), 141.
- (11) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (20), 12562.
- (12) Hamelberg, D.; McCammon, J. A. Fast peptidyl cis-trans isomerization within the flexible Gly-rich flaps of HIV-1 protease. *J. Am. Chem. Soc.* **2005**, *127* (40), 13778.
- (13) Markwick, P. R. L.; Bouvignies, G.; Salmon, L.; McCammon, J. A.; Nilges, M.; Blackledge, M. Toward a Unified Representation of Protein Structural Dynamics in Solution. *J. Am. Chem. Soc.* **2009**, *131* (46), 16968.
- (14) Markwick, P. R. L.; Cervantes, C. F.; Abel, B. L.; Komives, E. A.; Blackledge, M.; McCammon, J. A. Enhanced Conformational Space

Sampling Improves the Prediction of Chemical Shifts in Proteins. *J. Am. Chem. Soc.* **2010**, *132* (4), 1220.

(15) Grant, B. J.; Gorfe, A. A.; McCammon, J. A. Ras Conformational Switching: Simulating Nucleotide-Dependent Conformational Transitions with Accelerated Molecular Dynamics. *PLoS Comput. Biol.* **2009**, *5* (3), e1000325.

(16) Pierce, L. C. T.; Markwick, R. L.; McCammon, J. A.; Doltsinis, N. L. Accelerating Chemical Reactions: Exploring Reactive Free Energy Surfaces Using Accelerated Ab Initio Molecular Dynamics. *J. Chem. Phys.* **2010** submitted.

(17) Hamelberg, D.; de Oliveira, C. A. F.; McCammon, J. A. Sampling of slow diffusive conformational transitions with accelerated molecular dynamics. *J. Chem. Phys.* **2007**, *127* (15), 155102.

(18) Voter, A. F. Hyperdynamics: Accelerated molecular dynamics of infrequent events. *Phys. Rev. Lett.* **1997**, *78* (20), 3908.

(19) Car, R.; Parrinello, M. Unified Approach for Molecular-Dynamics and Density-Functional Theory. *Phys. Rev. Lett.* **1985**, *55* (22), 2471.

(20) Car Parrinello Molecular Dynamics (CPMD); IBM Corp.: Zurich, Switzerland, 2008; <http://www.cpmc.org/>.

(21) (a) Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic-Behavior. *Phys. Rev. A: At, Mol., Opt. Phys.* **1988**, *38* (6), 3098. (b) Lee, C. T.; Yang, W. T.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron-Density. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37* (2), 785.

(22) Martyna, G. J.; Tuckerman, M. E.; Tobias, D. J.; Klein, M. L. Explicit reversible integrators for extended systems dynamics. *Mol. Phys.* **1996**, *87* (5), 1117.

(23) Troullier, N.; Martins, J. L. Efficient Pseudopotentials for Plane-Wave Calculations. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1991**, *43* (3), 1993.

(24) Martyna, G. J.; Klein, M. L.; Tuckerman, M. Nose-Hoover Chains - the Canonical Ensemble Via Continuous Dynamics. *J. Chem. Phys.* **1992**, *97* (4), 2635.

(25) (a) Squillacote, M.; Sheridan, R. S.; Chapman, O. L.; Anet, F. A. L. Spectroscopic Detection of Twist-Boat Conformation of Cyclohexane - Direct Measurement of Free-Energy Difference between Chair and Twist-Boat. *J. Am. Chem. Soc.* **1975**, *97* (11), 3244. (b) Offenbach, J. L.; Fredin, L.; Strauss, H. L. Vibrational-Spectra of Twist-Boat Cyclohexane. *J. Am. Chem. Soc.* **1981**, *103* (5), 1001. (c) Gill, G.; Pawar, D. M.; Noe, E. A. Conformational study of cis-1,4-di-tert-butylcyclohexane by dynamic NMR spectroscopy and computational methods. Observation of chair and twist-boat conformations. *J. Am. Chem. Soc.* **2005**, *70* (26), 10726.

(26) Ross, B. D.; True, N. S. Nmr-Spectroscopy of Cyclohexane Gas-Phase Conformational Kinetics. *J. Am. Chem. Soc.* **1983**, *105* (15), 4871.

(27) Hura, G.; Sorenson, J. M.; Glaeser, R. M.; Head-Gordon, T. A high-quality x-ray scattering experiment on liquid water at ambient conditions. *J. Chem. Phys.* **2000**, *113* (20), 9140.

(28) Modig, K.; Pfrommer, B. G.; Halle, B. Temperature-dependent hydrogen-bond geometry in liquid water. *Phys. Rev. Lett.* **2003**, *90* (7), 075502-1.

(29) Eigen, M.; Demayer, L. Untersuchungen Uber Die Kinetik Der Neutralisation. *Z. Elektrochem.* **1955**, *59* (10), 986.

(30) Berkowitz, M.; Karim, O. A.; McCammon, J. A.; Rossky, P. J. Sodium-Chloride Ion-Pair Interaction in Water - Computer-Simulation. *Chem. Phys. Lett.* **1984**, *105* (6), 577.

DFT and Ab Initio Study of Iron-Oxo Porphyrins: May They Have a Low-Lying Iron(V)-Oxo Electromer?

Mariusz Radoń,^{*,†} Ewa Broclawik,[‡] and Kristine Pierloot[§]

[†]Faculty of Chemistry, Jagiellonian University, ul. Ingardena 3, 30-060 Kraków, Poland

[‡]Institute of Catalysis and Surface Chemistry, Polish Academy of Sciences, ul. Niezapominajek 8, 30-239 Kraków, Poland

[§]Department of Chemistry, University of Leuven, Celestijnenlaan 200F, B-3001 Heverlee-Leuven, Belgium

 Supporting Information

ABSTRACT: The energetics of various electromeric states for two heme complexes with an iron-oxo (FeO^{3+}) group, $\text{FeO}(\text{P})^+$ and $\text{FeO}(\text{P})\text{Cl}$ (P = porphin), have been investigated, employing DFT and correlated ab initio methods (CASPT2, RASPT2). Our interest focused in particular on tri- and pentaradicaloid iron(IV)-oxo porphyrin radical states as well as iron(V)-oxo states. Surprisingly, the iron(V)-oxo ground state is predicted for both models *in vacuo*. However, the presence of a polarizable medium, such as a solvent or a protein environment, favors the iron(IV)-oxo porphyrin radical cation, which is predicted to be the actual ground state of $\text{FeO}(\text{P})\text{Cl}$ under such conditions. Nonetheless, the iron(V)-oxo electromer is still expected to lie only a few kcal/mol above the ground state—a conclusion coming from both CASPT2 and RASPT2 calculations with a very large active space and further supported by a calibration with respect to coupled cluster CCSD(T) calculations for a simplified small model. The DFT results turn out to be strongly functional-dependent and thereby inconclusive. The widely used B3LYP functional—although correctly predicting the iron(IV)-oxo porphyrin radical ground state for $\text{FeO}(\text{P})\text{Cl}$ —seems to place the iron(V)-oxo states much too high in energy, as compared to the present CASPT2, RASPT2, and CCSD(T) results.

1. INTRODUCTION

High-valent iron-oxo porphyrins are strong and important oxidants, employed by nature in the cytochrome P450 catalytic cycle^{1,2} and supposedly participating in the oxygen-transfer reactions catalyzed by iron porphyrins.³ In both catalytic cycles, the active intermediate is normally assumed to be an iron(IV)-oxo porphyrin radical cation species, $(\text{Fe}^{\text{IV}}\text{O})(\text{P}^{\bullet+})$, also known as Compound I (Cpd I).¹ The high (+IV) oxidation state on Fe is stabilized by the coordination of strong electron donors: the porphyrinate and, particularly, the oxo ligand. In the presence of σ donors stronger than porphyrin (and more resistant to oxidation), such as tetradentate amido macrocycles, an even higher (+V) oxidation state of iron is stable in several known iron(V)-oxo complexes.^{4,5} The possibility of iron(V)-oxo complexes with corroles and corrolazines (close analogues of porphyrins) was also reported.⁶ All of these facts raise the important question of whether similar iron(V)-oxo species with a porphyrin ligand may also exist and be stable enough to be interesting for chemistry.^{7,8} Due to the very electrophilic nature of the $\text{Fe}^{\text{V}}\text{O}$ group (d^3), the hypothetical iron(V)-oxo porphyrin species could be extremely strong oxidants, with possibly very interesting properties.

The hypothetical iron(V)-oxo porphyrin may be viewed as an electromer (electronic isomer) of the “standard” Cpd I—i.e., iron(IV)-oxo porphyrin radical—species, obtained by moving one electron from the iron to the porphyrin, as illustrated in Figure 1. Furthermore, as shown in this figure, in both the iron(IV) and iron(V) forms, two local spin states on iron should be considered. For the iron(IV)-oxo states, these two possibilities are usually referred to as *triradicaloid* (local triplet state on the Fe,

denoted $^3\text{Fe}^{\text{IV}}$; three unpaired electrons in total) and *pentaradicaloid* (local quintet state on the Fe, denoted $^5\text{Fe}^{\text{IV}}$; five unpaired electrons in total). Similarly, for the iron(V)-oxo states, both the doublet and the quartet spin state are possible. The ligand radical in the triradicaloid and pentaradicaloid forms may be localized in either of the two highest-occupied π orbitals of the porphyrin ($\text{P}\pi$): a_{2u} or a_{1u} (the scheme in Figure 1 shows the first possibility).⁹ For each type of ligand radical, the coupling between the local spin on iron (triplet or quintet) and the local spin on the ligand (doublet) may be either ferro- or antiferromagnetic, thus producing a pair of close-lying electronic states: quartet and doublet (for the triradicaloid form) or sextet and quartet (for the pentaradicaloid form). Therefore, there are plenty of possible iron(IV) and iron(V) electronic states to be considered. Each of these states will be labeled by specifying the oxidation state and the local spin on the iron (e.g., $^3\text{Fe}^{\text{IV}}$, $^4\text{Fe}^{\text{V}}$) as well as the radical character and the local spin on the porphyrin: either closed-shell ^1P or the radical cation $^2\text{P}^{\bullet+}$ of two types (a_{2u} or a_{1u}); this notation is used in Figure 1.¹⁰

Various spectroscopic techniques have clearly identified the triradicaloid iron(IV)-oxo ground state for Cpd I species of enzymes and their synthetic analogues.^{11–16} A consensus has been reached between experiment and theory about this description.¹ Nonetheless, there have also been remarkable arguments in favor of iron(V)-oxo porphyrin species: from an early (and maybe naive) notion in the 1990s^{17,18} until the dissemination of recent spectroscopic data from laser flash photolysis (LFP)

Received: October 28, 2010

Published: March 23, 2011

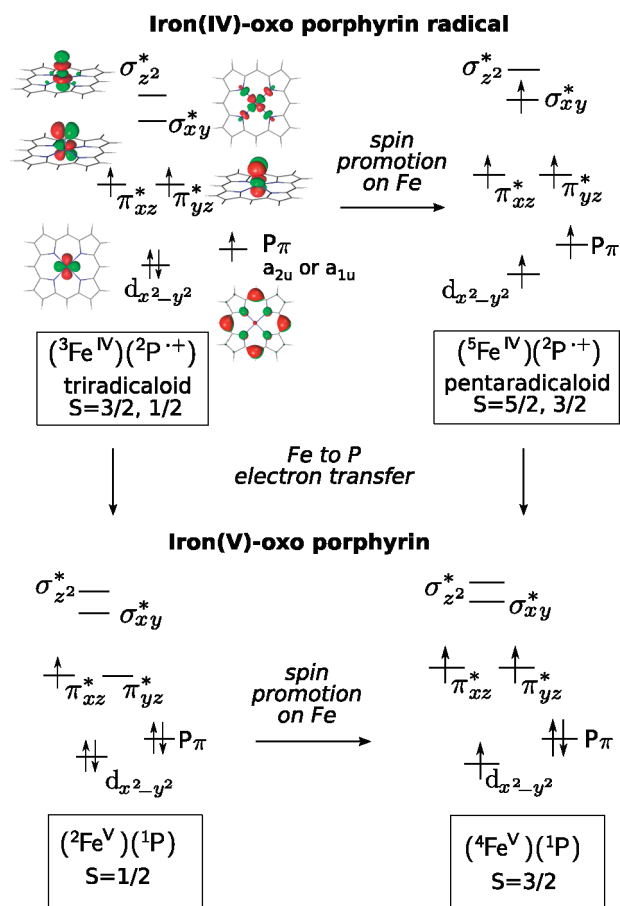


Figure 1. Electronic structure of an iron-oxo porphyrin compound in its various possible electromeric states.

experiments.^{19–21} The LFP experiments seem to suggest that the iron(V)-oxo electromer could be stable enough to be “seen” in UV/vis spectroscopy and kinetic experiments, and it was even speculated that the iron(V)-oxo species may play some role in oxo-transfer reactions catalyzed by iron porphyrins and enzymes.²¹ Yet, theory was rather sceptical of this concept. The iron(V)-oxo electromer of the P450 Cpd I model was indeed captured in some of the DFT studies, but at such a high energy (16–24 kcal/mol) above the iron(IV)-oxo triradicaloid ground state that it was judged as not likely to play any role in the catalytic oxidation pathways.^{22,23} Only very recently, Chen et al.²⁴ showed that with multireference ab initio methods (CASPT2) the iron(V)-oxo states appear at lower energies, suggesting that they might be readily accessible at ambient temperatures. Moreover, on the basis of recent theoretical studies, the iron(IV)-oxo pentaradicaloids were also proposed as alternative active species in the catalytic cycle.^{23,24} Despite notable progress, little is still known about the stabilities of various electromeric states in high-valent iron-oxo porphyrins (also in general, not necessarily restricted to models of P450 Cpd I). In our opinion, some aspects of the computational methodology also have to be further clarified, such as the performance of DFT and the adequacy of the active spaces used in the multireference calculations reported so far^{24–27} (in particular, the π system of porphyrin) in providing a realistic description of the various electromeric states. We note that a reliable theoretical description of Cpd I and related iron-oxo species really is a challenging task for all computational methods, having to

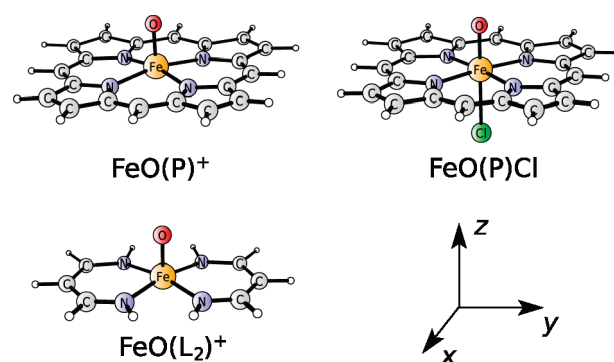


Figure 2. Structures of the studied heme complexes (FeO(P)^+ , FeO(P)Cl) and of the small model ($\text{FeO(L}_2\text{)}^+$), and their orientation in the coordinate system.

deal with at least two tricky issues: the problem of noninnocent ligands on the one hand and the description of spin state energetics in transition metal complexes on the other hand.⁸

In this paper, we study the relative energies of various electro-meric states in two model iron-oxo porphyrin complexes: FeO(P)^+ and FeO(P)Cl (P = porphyrin), shown in Figure 2. The complexes are studied using quantum-chemical calculations *in vacuo*, but the effects of a polarizable environment are also investigated. In addition to (now standard) DFT methods, we apply a multireference ab initio formalism: both the Complete Active Space (CASSCF/CASPT2) method²⁸ and its generalization—the Restricted Active Space (RASSCF/RASPT2) methods.²⁹ The RASPT2 method permits us to include more active orbitals in the active space than is possible in CASPT2, in particular, to extend the active space with a number of porphyrin π orbitals. On the basis of our previous experience with copper corroles³⁰ (also dealing with the question of metal versus ligand-based oxidation), the latter orbitals are expected to significantly influence the splitting between the iron(V)-oxo and iron(IV)-oxo porphyrin radical electromers. While considering various computational approaches for the present problem, we also thought about the coupled cluster CCSD(T) method. However, as the presently studied heme complexes are too large for this method, we applied CCSD(T) to a smaller model complex, $\text{FeO(L}_2\text{)}^+$ (where $\text{L} = \eta^2\text{-N}_2\text{C}_3\text{H}_5^-$ is the vinylogous amidine ligand), designed to mimic the essential structural and electronic features of the heme complexes, as shown in Figure 2. The idea of such a small model was inspired by similar “calibration” studies by Harvey et al. on ferrous and ferric heme systems.^{31,32} As we shall see, cross-checking of DFT, RASPT2, and CCSD(T) for the small model gives extra knowledge about the performance of these methods for the present task.

2. COMPUTATIONAL DETAILS

2.1. DFT Calculations. Spin-unrestricted DFT calculations were carried out with the Turbomole 5.9³³ and Gaussian 2009³⁴ packages, employing the def2-TZVP basis sets.³⁵ To cover a broad spectrum of various exchange-correlation functionals, we used both popular hybrid (B3LYP,³⁶ PBE0,³⁷ B3LYP*³⁸) and nonhybrid functionals (BP86,^{39,40} PBE,⁴¹ OLYP⁴²). We also tried the recently developed long-range corrected hybrid functional (LC- ω PBE^{43–46}), hoping that it might improve the description of electron transfer between the iron and porphyrin fragments. The structures were optimized at the BP86 level for

each of the considered electronic states (adiabatic calculations). DFT:BP86 structures were used in single point calculations with all other functionals (since initial tests showed that the relative energies are changing by just a fraction of a kcal/mol as compared to the energies obtained from full structure optimizations with each functional). The optimizations performed for the heme complexes exploited their C_{4v} symmetry, except for the degenerate ${}^2\text{Fe}^{\text{V}}$ doublet, which has an unequal occupation of the $\text{FeO } \pi_{xz,yz}^*$ orbitals and thus undergoes a significant Jahn–Teller distortion to C_{2v} . The distortion is more pronounced for $\text{FeO}(\text{P})^+$ than for $\text{FeO}(\text{P})\text{Cl}$. In order to preserve the similarity with the heme complexes, the small model was optimized with its four nitrogens constrained to lie in a single plane. The optimized Cartesian coordinates may be found in the Supporting Information. A number of DFT calculations including implicit solvation were also performed, making use of the standard COSMO model⁴⁷ as implemented in Turbomole. These calculations were performed for two values of the dielectric constant ($\epsilon = 5.7$ and 79), using the (unoptimized) bond radii multiplied by 1.17, and all other settings set as the defaults in Turbomole 5.9. Selected test calculations were also repeated with the PCM model⁴⁸ as implemented in Gaussian 2009. Very similar effects of solvation were obtained with both solvation models.

2.2. CASSCF/CASPT2 and RASSCF/RASPT2 Calculations. CASSCF/CASPT2²⁸ and RASSCF/RASPT2²⁹ calculations were performed with Molcas 7.4,⁴⁹ using a scalar-relativistic second-order Douglas–Kroll Hamiltonian,⁵⁰ the standard IPEA-shifted zero order Hamiltonian for second order perturbation theory,⁵¹ and Cholesky decomposition of two-electron repulsion integrals⁵²—all of these features as implemented in Molcas 7.4. Single-point calculations were performed on top of the DFT:BP86 structures obtained for each electronic state. The calculations for the heme complexes employed two types of Atomic Natural Orbitals (ANO) basis sets (basis I and II). The smaller one (basis I) was composed of ANO-RCC⁵³ on iron (contracted to $[7s6p5d2f1g]$) and ANO-S⁵⁴ on the ligands (contracted to $[4s3p1d]$ on C, N, and O, to $[5s4p2d]$ on Cl, and to $[2s]$ on H). The larger one (basis II) was composed of ANO-RCC on all atoms, contracted to $[7s6p5d3f2g1h]$ on Fe; to $[4s3p2d1f]$ on C, N, and O; to $[5s4p3d2f]$ on Cl; and to $[3s1p]$ on H. The results reported in section 3 were obtained with basis II, while basis I was used mostly for testing purposes (see the Supporting Information). The calculations for the small model $\text{FeO}(\text{L}_2)^+$ additionally employed three types of correlation consistent basis sets, the same as used in the CCSD(T) calculations and denoted as T/D, T/T, and Q/T (*vide infra*). In all CASPT2 and RASPT2 calculations, the core electrons were kept frozen. However, for bases I and II, the semicore Fe ($3s$, $3p$) electrons were correlated, in contrast to the correlation consistent basis sets (T/D, T/T, Q/T), which are not designed for correlating these electrons. It was tested (with basis I) that excluding the Fe ($3s$, $3p$) electrons from the correlation treatment only gives a secondary difference in the relative energies (affecting mostly the energy difference between the pentaradicaloid and triradicaloid Fe^{IV} states, while leaving the separation between the Fe^{V} and Fe^{IV} states virtually unaffected).

All of the CASSCF/CASPT2 and RASSCF/RASPT2 calculations were performed state specifically for each of the considered electronic states. The high symmetry of the present models (C_{4v} for the heme models, C_{2v} for the small model) is beneficial for locating the interesting states as the lowest CI roots corresponding to a given spin and spatial symmetry. Actually, due to

technical limitations (Molcas supports Abelian groups only), all of the calculations were performed in C_{2v} formal symmetry. Most of the considered electronic states were still located as the lowest roots in different irreps of C_{2v} . This is not the case with the ${}^4\text{A}_2$ and ${}^4\text{B}_2$ states, both belonging to the same irrep (A_2) in C_{2v} . However, as they are still orthogonal, the ${}^4\text{A}_2$ and ${}^4\text{B}_2$ states were easily distinguished from each other in state-specific calculations (with help of the CISELECT facility of the RASSCF module in Molcas). Obviously, it would be more difficult to apply the same state-specific procedure to less symmetric systems (not covered in this study), like for instance P450 Cpd I. In general, symmetry lowering may cause mixing between the Fe^{IV} and the Fe^{V} states, analogous that found here for the ${}^2\text{B}_1$ state of the small model (*vide infra*).

The choice of the active orbitals for the CASSCF calculations was made according to the standard rules for transition metal compounds.^{55–57} Nondynamical correlation effects involving the Fe $3d$ electrons, the Fe–O bond, and the Fe–P σ bond are described by making active four pairs of bonding–antibonding orbitals ($\text{Fe}3d_{xz}-\text{O}2p_x \rightarrow (\pi_{xz}, \pi_{xz}^*)$; $\text{Fe}3d_{yz}-\text{O}2p_y \rightarrow (\pi_{yz}, \pi_{yz}^*)$; $\text{Fe}3d_{z^2}-\text{O}2p_z \rightarrow (\sigma_{z^2}, \sigma_{z^2}^*)$; $\text{Fe}3d_{xy}-\text{P}\sigma_{xy} \rightarrow (\sigma_{xy}, \sigma_{xy}^*)$), the remaining nonbonding Fe $3d_{x^2-y^2}$ orbital, and three double-shell orbitals ($3d'_{xz}$, $3d'_{yz}$, $3d'_{x^2-y^2}$). In order to allow electron transitions between the Fe and P fragments, the two highest-occupied π orbitals of the porphyrin, $a_{2u}(a_1)$ and $a_{1u}(a_2)$, as well as their correlating π^* orbitals $e_g(b_1, b_2)$ were made active. These four frontier π orbitals of the porphyrin are known in the literature as the *Gouterman set*. This choice leads to an active space of 15 electrons distributed in 16 active orbitals (15in16). Although this is already a fairly large active space, previous experience⁵⁸ tells us that many more (preferably all) π , π^* orbitals on porphyrin should be made active in order to provide a correct description of the relative energies of various electronic states (in particular, the Fe^{V} and Fe^{IV} states). Obviously, this is computationally not feasible within CASSCF. Therefore, RASSCF calculations were performed instead. Here, the active space is further subdivided into three subspaces, RAS1, RAS2, and RAS3.²⁹ The RAS2 subspace (where, as in CASSCF, all possible excitations are allowed) was kept limited to the singly occupied orbitals plus those pairs of orbitals describing the most important nondynamical correlation effects. In contrast, the doubly occupied π orbitals of the porphyrin and their correlating π^* orbitals, as well as other less important active orbitals (among which the three $3d'$ orbitals), were kept in RAS1 (if nearly doubly occupied) or in RAS3 (if nearly empty). Up to double excitations were then allowed out of RAS1 and into RAS3. Similar calculations have already been performed recently, both for organic molecules (such as free base porphyrin) and transition metal systems (such as copper corrole).^{29,30,58} However, for the present case, it turned out to be particularly difficult to find a combination of a global active space with an RAS2 subspace that is both reliable and still computationally feasible. A number of test calculations have been carried out for this purpose: selected results are given in the Supporting Information, and the most important conclusions obtained from these preliminary studies are summarized below.

First, extending the 15in16 active space with all remaining π and π^* orbitals of the porphyrin leads to a very large active space of 37 electrons in 36 orbitals (37in36). This active space is computationally feasible only if combined with a very small RAS2 subspace, containing only the singly occupied (SO) orbitals. Therefore, 37in36 was reduced by removing four occupied π orbitals of the porphyrin and their four correlating π^* orbitals

(eight orbitals in total). The orbitals removed correspond to combinations of $C2p_z$ of the β carbons of the pyrrole rings. The resulting active space of 29 electrons in 28 orbitals (29in28) still contains 16 (π, π^*) orbitals on the porphyrin and may be viewed as an active space correlating 18 “aromatic electrons” of the porphyrin ring (Hückel formula: $4n + 2 = 18$ for $n = 4$, corresponding to the shortest cyclic path through the porphyrin ring). A comparison of the results obtained with the 37in36 and 29in28 active spaces, combined with a RAS2 subspace containing only the singly occupied orbitals, has shown that both active spaces point to a similar (up to ± 3 kcal/mol) set of relative energies (see the Supporting Information).

The second step was to find a reasonable RAS2 subspace to be combined with the 29in28 global active space. To this end, various choices of RAS2 were compared within the smaller 15in16 global active space (see the Supporting Information). First, it turns out that a RASPT2(15in16) treatment based on a RAS2(15in11) subspace (i.e., with the Fe 3d' and the correlating π^* (e_g) orbitals on the porphyrin moved to RAS3 as compared to CAS) produces nearly the same relative energies as the full CASPT2(15in16) treatment. This brings a promising suggestion that both the 3d double-shell effect on iron and the porphyrin $\pi-\pi^*$ correlation may be well described by up to double excitations. Next, RASPT2(15in16) calculations were tested in combination with two smaller subspaces, denoted as RAS2(6+SO) and RAS2(8+SO). RAS2(6+SO) contains three couples of bonding–antibonding Fe3d–O2p orbitals and all (remaining) singly occupied orbitals (i.e., 7in6 for 2E , 9in7 for $^4A_{1,2}$, 11in9 for $^6A_{1,2}$). With this RAS2 space, an optimal description of the non-dynamical correlation effects connected to the (very covalent) Fe–O bond is provided. RAS2(8+SO) additionally contains the $\sigma_{xy}\sigma_{xy}^*$ pair, thus improving the description of nondynamical correlation related to σ donation from P to Fe. Obviously, the RASPT2 results obtained with RAS2(6+SO) and RAS2(8+SO) differ somewhat from each other and from those obtained with RAS2(15in11), as well as from the full CASPT2 results. However, the differences are rather small (a few kcal/mol), indicating that even the least extensive subspace RAS2(6+SO) is a reasonable choice. The results reported below refer to the 29in28 global active space combined with the RAS2(6+SO) subspace. The choice of RAS2(8+SO) would perhaps give slightly better results. However, this subspace is too large to be combined with the 29in28 global space at a reasonable computational cost. The small variation of the relative energies caused by excluding the $\sigma_{xy}\sigma_{xy}^*$ pair from RAS2 in RAS2(6+SO) is not expected to change any of our important conclusions below.

The active space for the small model was constructed analogously. Since we have not found any natural analogue of the Goutermann π set for the small model, the 15in16 active space was not considered, and hence no CASSCF/CASPT2 calculations were performed for the small model. Including all ligand π and π^* orbitals gives an active space of 23 electrons in 22 orbitals (23in22) for RASSCF/RASPT2 calculations. This active space is still small enough to be combined with various choices of RAS2 (defined analogously as for the heme models): RAS2(6+SO), RAS2(8+SO), and RAS2(15in11). (The last subspace contains the highest occupied π orbitals of a_1 and b_2 symmetries instead of a_{2u} and a_{1u} for porphyrin.) The orbital carrying the free radical in the Fe^{IV} states considered for the small model, HOMO(b_2), was treated as singly occupied not only for the 4B_1 and 6B_2 states but also for 2B_1 , where it is partially occupied due to the mixed Fe^V–Fe^{IV} character of this state (*vide infra*). In contrast, this

orbital was treated as doubly occupied for the 2B_2 state. Our main conclusions drawn from these test calculations essentially confirm the picture obtained for the heme complexes. Thus, the 23in22 results obtained with RAS2(15in11) and RAS2(8+SO) are nearly identical and also similar to those obtained with the RAS2(6+SO) subspace. The largest discrepancies between the RAS2(6+SO) and RAS2(8+SO) subspaces are observed for the tri- to pentaradicaloid excitation energy. However, we shall see that the CASPT2/RASPT2 description of this excitation is almost certainly biased in favor of the high-spin state anyway. In contrast, the difference between the predicted Fe^V–Fe^{IV} gaps with both RAS2 spaces is much less pronounced. The results reported for the small model in section 3 refer to the 29in28 global active space with a RAS2(8+SO) subspace. More technical details of CASSCF/CASPT2 and RASSCF/CASPT2 calculations may be found in the Supporting Information.

2.3. CCSD(T) Calculations for the Small Model. Restricted open-shell CCSD(T) calculations were performed for DFT: BP86 structures with Molpro 2009⁵⁹ using the Douglas–Kroll (DK) scalar relativistic Hamiltonian and three DK reconstructions⁶⁰ of the Dunning correlation consistent (cc) basis sets.⁶¹ The smallest one, denoted T/D, was composed of cc-pVTZ-DK on Fe and cc-pVDZ-DK on the ligands. The middle one, T/T, was simply cc-pVTZ-DK on all atoms. The most extensive one, Q/T, was composed of cc-pVQZ-DK on Fe and cc-pVTZ-DK on the ligands. The correlation energy extrapolated to the complete basis set on Fe (while keeping cc-pVTZ-DK on the ligands) was also calculated from the Q/T and T/T results assuming the “1/ X^3 ” dependence of the residual correlation energy due to Helgaker et al.⁶²—yielding the results reported below as ∞ /T. The restricted open-shell Kohn–Sham orbitals obtained from the B3LYP functional were used in the CCSD(T) calculations in order to speed up the coupled cluster convergence and to reduce the orbital bias caused by electron correlation (for the use of Kohn–Sham orbitals in coupled cluster calculations, we refer to the literature^{32,63–65}). The impact of multireference effects on the coupled cluster calculations was estimated by means of various diagnostics computed for the CCSD wave function: \mathcal{F}_1 ,⁶⁶ \mathcal{D}_1 ,⁶⁷ and the maximum absolute value found for the amplitudes of double-excitations ($\max_{ij,ab} |t_{ij}^{ab}|$).

3. RESULTS AND DISCUSSION

3.1. Electromeric States of FeO(P)⁺ and FeO(P)Cl. Before discussing the stability of various electromeric forms for the heme models, FeO(P)⁺ and FeO(P)Cl, let us first describe which of their electronic states are included in this study. The various electromeric forms (cf Figure 1) give rise to the following electronic states (all labeled under C_{4v} symmetry):

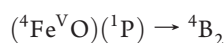
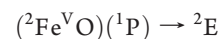
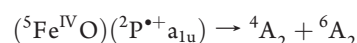
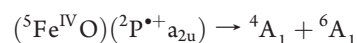
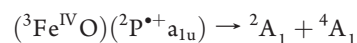
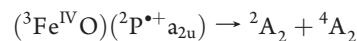


Table 1. Relative Energies (kcal/mol) of the Low Lying Electronic States of FeO(P)⁺a

	B3LYP	B3LYP*	OLYP	BP86	CASPT2 ^b	RASPT2 ^c
(³Fe ^{IV} O)(²P ⁺ a _{2u}) ⁴A ₂	0	0	0	0	0	0
	²A ₂	0.4	0.4	0.4	0.4	-0.6
(³Fe ^{IV} O)(²P ⁺ a _{1u}) ⁴A ₁	-3.3	-2.8	-1.2	-1.5	6.4	-0.7
	²A ₁	-3.4	-2.9	-1.1	-1.3	5.6
(⁵Fe ^{IV} O)(²P ⁺ a _{2u}) ⁶A ₁	8.9	12.4	9.7	19.1	3.9	0.1
(⁵Fe ^{IV} O)(²P ⁺ a _{1u}) ⁶A ₂	5.7	9.9	9.1	18.3	11.2	-0.8
(²Fe ^V O)(¹P) ²E	12.7	9.2	0.5	-2.4	-3.4	-6.5
(⁴Fe ^V O)(¹P) ⁴B ₂	12.8	11.3	3.4	5.7	-1.6	-3.5

^a CASPT2 and RASPT2 energies obtained with basis II, DFT energies with def2-TZVP (see section 2). ^b Based on the 15in16 active space, including only the four π (Gouterman) orbitals on the porphyrin. ^c Based on the 29in28 active space, including 16 π orbitals on the porphyrin, with RAS2(6+SO).

Table 2. Relative Energies (kcal/mol) of the Low Lying Electronic States of FeO(P)(Cl)^a

	B3LYP	B3LYP*	OLYP	BP86	CASPT2 ^b	RASPT2 ^c
(³Fe ^{IV} O)(²P ⁺ a _{2u}) ⁴A ₂	0	0	0	0	0	0
	²A ₂	0.1	0.1	0.1	0.1	-1.4
(³Fe ^{IV} O)(²P ⁺ a _{1u}) ⁴A ₁	6.3	6.8	8.0	8.1	16.5	9.1
	²A ₁	5.8	6.4	7.8	7.9	15.5
(⁵Fe ^{IV} O)(²P ⁺ a _{2u}) ⁶A ₁	9.7	12.9	10.1	19.1	3.2	-1.5
(⁵Fe ^{IV} O)(²P ⁺ a _{1u}) ⁶A ₂	17.1	20.7	18.7	27.8	21.1	9.6
(²Fe ^V O)(¹P) ²E	12.4	13.3	2.9	1.0	1.6	-1.7
(⁴Fe ^V O)(¹P) ⁴B ₂	13.1	11.5	3.5	5.9	3.1	2.8

^a CASPT2 and RASPT2 energies obtained with basis II, DFT energies with def2-TZVP (see section 2). ^b Based on the 15in16 active space, including only the four π (Gouterman) orbitals on the porphyrin. ^c Based on the 29in28 active space, including 16 π orbitals on the porphyrin, with RAS2(6+SO).

Due to the high symmetry of the heme complexes, mixing between most of the listed electronic states is symmetry-forbidden. Exceptions are the quartets of A_{1,2} symmetry, each emerging both from the triradicaloid and the pentaradicaloid form. However, because of the different local spin on the iron for both forms, virtually no mixing is observed in this case. We further note that for the pentaradicaloid form (⁵Fe^{IV})(²P⁺), only the sextet states will be reported below (the corresponding quartet states are expected to have very similar energies²⁴).

Tables 1 and 2 contain the relative (adiabatic) energies obtained for the considered states of the studied heme complexes, FeO(P)⁺ and FeO(P)Cl; in all cases, the energies are given with respect to ⁴A₂ (i.e., the triradicaloid with a_{2u} hole).

Let us first focus on the iron(IV)-oxo triradicaloid states. As might be expected, the doublet and the quartet states of each triradicaloid (a_{2u}- or a_{1u}-type) are close in energy—a situation which is also known well for the models of P450 Cpd I.^{1,68} Looking at the relative energies of the states with either a_{2u} or a_{1u} orbitals singly occupied (i.e., considering the ⁴A₂–⁴A₁ or ²A₂–²A₁ gap), one may easily notice that the different DFT methods predict rather similar relative energies, which are also close to the RASPT2 results. In contrast, the CASPT2 calculations significantly overstabilize the a_{2u}-type radicals. This is obviously caused by the limitations of the 15in16 active space

underlying the CASPT2 calculations; this active space includes only four frontier π orbitals on the porphyrin (the Gouterman set). The deficiency is clearly removed in the RASPT2 calculations based on the larger 29in28 active space, containing now as many as 16 π orbitals on the porphyrin. We further notice that the a_{1u}-type radicals (²⁴A₁) are more stable than the a_{2u}-type radicals (²⁴A₂) for FeO(P)⁺ (except for the deficient CASPT2 calculations), whereas in FeO(P)Cl, this ordering is reversed. It was noted just above that within the same configuration (either a_{1u} or a_{2u} radical), a very small quartet–doublet energy separation is found at all computational levels. This is easily understandable, as both states differ only by weak magnetic coupling between two spatially separated spins: on the FeO group and on the porphyrin. However, from the results in Tables 1 and 2, it is clear that also for this property CASPT2 gives results that are deviating both from the DFT and the RASPT2 results, by overestimating the absolute value of the magnetic coupling and often giving it a wrong sign (i.e., predicting antiferromagnetic instead of ferromagnetic coupling). Again, these deviating results should be brought back to the limitations of the 15in16 active space used in CASPT2. Using a larger active space, the RASPT2 calculations clearly provide results for the ²A₂–⁴A₂ and ²A₁–⁴A₁ gaps that are superior to those of CASPT2.

Let us now focus on the energy difference between the pentaradicaloid (⁵Fe^{IV})(²P⁺) and the corresponding triradicaloid (³Fe^{IV})(²P⁺) states. As was already mentioned (and is shown in Figure 1), the transition from the tri- to pentaradicaloid state *de facto* comes down to a spin promotion from the intermediate-spin (triplet) to the high-spin (quintet) local spin state of Fe^{IV}. Due to the local character of this excitation, about the same energy difference is obtained for the a_{2u}-type radicals (i.e., ⁶A₁–⁴A₂ gap) as for the a_{1u}-type radicals (i.e., ⁶A₂–⁴A₁ gap). It is known from previous studies on transition metal complexes^{8,69–71} that spin state energetics are very sensitive to the applied DFT method. This is also the case here, the hybrid functionals predicting a much lower relative energy of the sextet (spin-promoted) state than the nonhybrid functionals. The discrepancies are substantial, for instance BP86 predicts the ⁶A₁ state at ~19 kcal/mol above the ⁴A₂ state, while B3LYP gives only ~9–10 kcal/mol. Considering the spin promotion energy, the OLYP (nonhybrid) functional performs closer to the hybrid than to other nonhybrid functionals, such as BP86 or PBE (the latter results are given in the Supporting Information). We next observe that the CASPT2 and RASPT2 calculations clearly favor the high-spin states, predicting even smaller promotion energies than those from the hybrid functionals. A similar result was also obtained by Chen et al. in their CASPT2/MM study of P450 Cpd I.²⁴ However, recent studies on a number of iron(II) complexes in N₄ and N₂O₂ architectures^{72–74} have indicated that CASPT2 has a tendency to overstabilize the high spin (quintet) state in these systems with at least a few kcal/mol. All problematic cases involve an electron excitation from the nonbonding iron 3d orbital to the antibonding iron–ligand combination, d_{x²-y²} → σ_{xy}^* . The same type of electron excitation is also involved here. We therefore suspect that the present CASPT2 results may be biased toward the pentaradicaloid states by a few kcal/mol, a problem that is clearly not solved by going to RASPT2. Further confirmation of this observation will be provided by comparing RASPT2 with CCSD(T) benchmark calculations for the small model (*vide infra*).

Perhaps the most exciting property of the studied complexes is the relative stability of the iron(V)-oxo and the iron(IV)-oxo

porphyrin radical electromers. As observed previously for the spin promotion energy, here also the DFT results turn out to be substantially functional-dependent. The hybrid functionals, like B3LYP, B3LYP*, and PBE0 (the latter results are given in the Supporting Information), strongly favor the iron(IV)-oxo triradicaloid states (${}^2,{}^4A_1$ or ${}^2,{}^4A_2$) and predict a very high energy for the iron(V)-oxo states (2E , 4B_2). In contrast, the nonhybrid functionals, like BP86, OLYP, and PBE, place both types of states within a few kcal/mol margin. Actually, BP86 and PBE predict the 2E state as the ground state in $FeO(P)^+$ and as nearly degenerate with the (triradicaloid) ground state in $FeO(P)Cl$. Reducing the amount of HF exchange from 20 to 15% in the B3LYP* functional does bring the predictions of this functional somewhat closer to the nonhybrid functionals than is the case for B3LYP (although not always—see the 2E state in $FeO(P)Cl$), but still the discrepancy is huge. We also applied the more recent, range-separated hybrid functional LC- ω PBE, hoping that it might improve the description of electron transfer between the iron and the porphyrin.⁴⁴ However, for the presently studied systems, this functional performs very similarly to the traditional hybrid functionals (the results may be found in the Supporting Information). In sum, we have found that the hybrid and nonhybrid functionals point to two completely different pictures, with differences in relative energies sometimes much larger than 10 kcal/mol. Which of these two pictures is then closer to reality?

The multireference RASPT2(29in28) and CASPT2(15in16) calculations rather unambiguously place the iron(V)-oxo states at low energy. This is qualitatively similar to the predictions from the nonhybrid functionals. Such a low energy for the iron(V)-oxo electromer may seem rather surprising. In fact, after performing the CASPT2(15in16) calculations, we initially suspected that the iron(V)-oxo states might be overstabilized by this method, given the limitations of the 15in16 active space. This was our primary motivation to introduce the extended 29in28 active space, by including more π, π^* orbitals on the porphyrin, and thereby improving the description of the electron transfer from the iron to the macrocycle. To our surprise, this enlargement of the active space turned out to stabilize the iron(V)-oxo states even more! In fact, the RASPT2 calculations predict an iron(V)-oxo ground state (2E) for both complexes. This is the situation predicted for isolated molecules *in vacuo*; as will be shown below, the presence of solvent or another polarizable medium is important for the relative energetics and may even change the ground state. We note that in a recent CASPT2/MM study by Chen et al.,²⁴ the iron(V)-oxo electromer was also found at low energy for Cpd I of cytochrome P450. The authors of the cited paper used an active space analogous to our 15in16 space, though even slightly smaller on the porphyrin. We expect that extension of their active space (i.e., providing a more balanced description of charge transfer between the porphyrin and the FeO group) would even further stabilize the iron(V)-oxo electromer of Cpd I.

In previous DFT studies, the iron(V)-oxo states of P450 Cpd I were reported as electronically unstable in the presence of a polarizable continuum or of point charges (i.e., the SCF procedure under such conditions converging to the “usual” triradicaloid state).^{2,22,23} Here, we estimated the effect of polarization on the relative energetics of various electromers by means of the standard COSMO model.⁴⁷ The calculations with a dielectric continuum obviously cannot provide a quantitative description of the $FeO(P)^+$ and $FeO(P)Cl$ species in solution—their only purpose is to provide a qualitative estimation of (nonspecific)

Table 3. Effect of the Polarizable Continuum (COSMO) on the Relative Energetics in $FeO(P)^+$ (kcal/mol)

		$\epsilon = 5.7$		$\epsilon = 78$	
		B3LYP	BP86	B3LYP	BP86
$({}^3Fe^{IV}O)({}^2P^{*+}_{a_{2u}})$	4A_2	0	0	0	0
$({}^3Fe^{IV}O)({}^2P^{*+}_{a_{1u}})$	4A_1	0.0	0.2	0.0	0.2
$({}^5Fe^{IV}O)({}^2P^{*+}_{a_{2u}})$	6A_1	0.6	0.4	0.8	0.6
$({}^5Fe^{IV}O)({}^2P^{*+}_{a_{1u}})$	6A_2	0.2	0.3	0.2	0.3
$({}^2Fe^VO)({}^1P)$	2E	2.7	2.3	3.7	3.2
$({}^4Fe^VO)({}^1P)$	2B_2	1.5	1.5	2.2	2.1

Table 4. Effect of the Polarizable Continuum (COSMO) on the Relative Energetics in $FeO(P)Cl$ (kcal/mol)

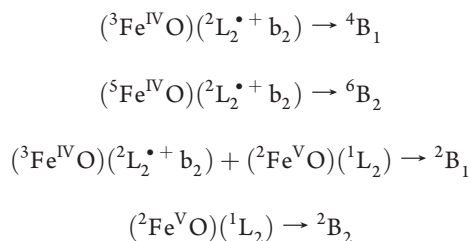
		$\epsilon = 5.7$		$\epsilon = 78$	
		B3LYP	BP86	B3LYP	BP86
$({}^3Fe^{IV}O)({}^2P^{*+}_{a_{2u}})$	4A_2	0	0	0	0
$({}^3Fe^{IV}O)({}^2P^{*+}_{a_{1u}})$	4A_1	−1.7	−1.9	−2.2	−2.5
$({}^5Fe^{IV}O)({}^2P^{*+}_{a_{2u}})$	6A_1	1.1	0.9	1.6	1.3
$({}^5Fe^{IV}O)({}^2P^{*+}_{a_{1u}})$	6A_2	0.0	−0.4	0.1	−0.3
$({}^2Fe^VO)({}^1P)$	2E	5.7	5.6	7.9	7.9
$({}^4Fe^VO)({}^1P)$	2B_2	4.4	3.4	6.2	4.9

environmental effects in order to distinguish the situation *in vacuo* from the situation in solution or in a protein. The calculations employed two values for the dielectric constant: $\epsilon = 5.7$ (corresponding to a weakly polar solvent such as chlorobenzene and often used for modeling the weakly polar interior of proteins) and $\epsilon = 79$ (corresponding to a polar solvent—water) and were performed at the DFT level with two functionals: B3LYP and BP86. The estimated effects of the polarizable continuum on the relative energetics are given in Tables 3 and 4. The numbers provided in these tables are the differences in relative energy, with respect to the 4A_2 state, between the COSMO and the vacuum calculations.

First, it can be seen that both functionals—even though predicting very different relative energetics (cf. Tables 1 and 2)—now point to quite similar “solvent” effects. This was to be expected, as the effect of a polarizable continuum is due to simple electrostatics, in contrast to the ordering of various electromeric states, which is heavily dependent on electron correlation. Second, the primary effect of the polarizable continuum is to destabilize the iron(V)-oxo states with respect to the iron(IV)-oxo states. The effect is more pronounced for $FeO(P)Cl$ and is essentially achieved already at the smaller value of the dielectric constant ($\epsilon = 5.7$). Assuming that the “solvent” effect calculated at the DFT level is transferable to RASPT2 energetics, one would obtain the iron(IV)-oxo electromer as the ground state in $FeO(P)Cl$. The iron(V)-oxo electromer is predicted to be an excited state lying at only ~ 4 – 6 kcal/mol above (depending on ϵ). In $FeO(P)^+$, the iron(V)-oxo electromer is still predicted to be the ground state, but the excited states of iron(IV)-oxo porphyrin radical character are expected at lower energies than *in vacuo*. In sum, the polarizable medium appears to be an important factor in determining the relative energetics of iron(IV)-oxo porphyrin radical and iron(V)-oxo electromers.

3.2. Cross-Checking the Methods for the Small Model. The studies on $\text{FeO}(\text{P})^+$ and $\text{FeO}(\text{P})\text{Cl}$ have clearly demonstrated how much the calculated stabilities of various electromers are sensitive to the choice of methodology. Looking for further assessment of our computational methods, we focused on the small model, $\text{FeO}(\text{L}_2)^+$ ($\text{L} = \eta^2\text{-N}_2\text{C}_3\text{H}_5^-$). This model mimics the basic electronic features of the heme complexes but is, on the other hand, small enough to compare the results obtained from DFT and RASPT2 calculations with the CCSD(T) method. From an inspection of the various diagnostics of multireference effects in the coupled cluster wave function (see section 2 and the Supporting Information),^{31,32,75,76} we conclude that the small model is still a well-behaving case for the CCSD(T) method. Therefore—as we have also taken care regarding the use of large enough basis sets—we believe that the CCSD(T) results given below may be considered reliable reference data.

Before discussing the stabilities of various electromeric states in the small model, let us first give an overview of the electronic states considered in the present study. The electronic situation of the small model is, obviously, slightly different from the heme complexes, because of the smaller size of the π system of L_2 as well as the lower symmetry. As for possible $(\text{Fe}^{\text{IV}}\text{O})(\text{L}_2^{\bullet+})$ states, only those corresponding to the radical in the highest occupied b_2 π orbital of the ligand ($\text{L}_2^{\bullet+}b_2$) were calculated. Due to the lower (C_{2v}) symmetry, the $\text{Fe}^{\text{V}}\text{O}$ doublet is no longer degenerate. Moreover, one of its components, ${}^2\text{B}_1$, strongly mixes with the $(\text{Fe}^{\text{IV}}\text{O})(\text{L}_2^{\bullet+})$ doublet configuration of the same symmetry, thus producing a ${}^2\text{B}_1$ electronic state of mixed ($\text{Fe}^{\text{V}} + \text{Fe}^{\text{IV}}$) nature. In contrast, the other $\text{Fe}^{\text{V}}\text{O}$ doublet state, ${}^2\text{B}_2$, remains predominantly iron(V)-oxo in character. The RASSCF calculations for the ${}^4\text{Fe}^{\text{V}}$ state (here, ${}^4\text{A}_2$) using the same active space as for the other electronic states were not successful due to unfavorable orbital rotations; for this reason, the result for this state will not be reported below (anyway, we believe that the results obtained for the other states already carry substantial information). In sum, the following electronic states have been studied for the small model:



Selected results obtained from the DFT, RASPT2, and CCSD(T) calculations are given in Table 5 (as for the heme complexes, all energies are given with respect to the triradicaloid state, ${}^4\text{B}_1$). The table includes two sets of RASPT2(23in22) results (the active space covering all π, π^* orbitals of L_2), obtained with two different basis sets: basis set II (i.e., the one used for the heme complexes in Tables 1 and 2) and the basis set Q/T (i.e., the one used in the CCSD(T) calculations). We note that both sets of RASPT2 energies agree very well. CCSD(T) relative energies are given for the Q/T basis as well as extrapolated to an infinite basis set on iron (∞/T). Also here, both sets of results are reasonably similar, suggesting that they are already converged with respect to the basis set.

Looking at Table 5, one should not forget that the bis-(vinylous amidine) ligand (L_2^{2-}) is obviously different from porphyrinate (P^{2-})—therefore, the stabilities of various electromeric

Table 5. Relative Energies (kcal/mol) for the Low-Lying Electronic States of the Small Model, $\text{FeO}(\text{L}_2)^+$

	RASPT2 ^a				CCSD(T)			
	B3LYP	B3LYP*	OLYP	BP86	II ^b	Q/T ^b	Q/T ^b	∞/T^b
$({}^3\text{Fe}^{\text{IV}})(\text{L}_2^{\bullet+}) {}^4\text{B}_1$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$({}^5\text{Fe}^{\text{IV}})(\text{L}_2^{\bullet+}) {}^6\text{B}_2$	17.0	20.6	17.6	28.0	11.2	12.1	16.3	17.8
mixed ^c ${}^2\text{B}_1$	1.5	-1.5	-8.6	-10.4	-12.2	-12.0	-8.9	-8.5
$({}^2\text{Fe}^{\text{V}})({}^1\text{L}_2) {}^2\text{B}_2$	8.8	5.6	-0.5	-2.7	-1.7	-2.1	-4.2	-3.7

^a RAS(23in22) with RAS2(8+SO). ^b See the Supporting Information for a definition of the basis sets. ^c A mixture of doublet ($({}^3\text{Fe}^{\text{IV}})(\text{L}_2^{\bullet+})$) and ($({}^2\text{Fe}^{\text{V}})({}^1\text{L}_2)$) configurations.

states are also clearly different from those of the heme complexes. Nevertheless, the *trends* observed previously for the relative energies remain virtually the same. For instance, the nonhybrid functionals and RASPT2 predict the Fe^{V} (${}^2\text{B}_2$) at much lower relative energy than the hybrid functionals. The same is also observed for the state of mixed $\text{Fe}^{\text{V}}-\text{Fe}^{\text{IV}}$ character (${}^2\text{B}_1$). The difference in the relative energies of ~ 10 kcal/mol between the B3LYP and BP86 is almost quantitatively comparable to that found previously for the heme models (cf. Tables 1 and 2). Moreover, the spin promotion energy (${}^6\text{B}_2-{}^4\text{B}_1$) is much higher according to BP86 than according to B3LYP, B3LYP*, OLYP, and RASPT2. This is, again, in nearly quantitative agreement with the situation of the heme complexes. Thus, the overall similarity of the small model to the heme complexes justifies our choice of this system for testing various computational methodologies. It also suggests that the electronic structure issues discussed here and causing the discrepancies between different functionals are quite general features of high-valent iron-oxo complexes with π macrocycles. They are, apparently, not limited only to the special case of porphyrin.

When comparing the CCSD(T) and RASPT2 results, two observations may be made. First, RASPT2 most likely underestimates the spin promotion energy (overstabilizes the high-spin state) by some 4–6 kcal/mol, as was in fact expected (*vide supra*). This bias of RASPT2 in favor of the high-spin state is analogous to the one found previously in CASPT2 calculations of similar $d_{xz-yz} \rightarrow \sigma_{xy}^*$ spin-promotions in iron(II) complexes with N_4 and N_2O_2 ligands (e.g., heme, salen).^{72–74} (N.B., also there, CCSD(T) calculations for small models were helpful to diagnose this effect.) For the present case, the spin promotion energy seems to be most correctly described by the B3LYP and OLYP functionals. The second important observation is that RASPT2 and CCSD(T) closely agree on the relative stability of the Fe^{V} and Fe^{IV} electromers. This is positive, providing extra credibility to this aspect of the RASPT2 energetics for the heme complexes. Thus, the CCSD(T) calculations for the small model indirectly support the previous conclusion about the low-lying iron(V)-oxo electromer in heme complexes.

3.3. Comparison with Experiment. As was already mentioned in the Introduction, the iron(IV)-oxo triradicaloid ground state is experimentally confirmed for enzymatic Cpd I species and their analogues, on the basis of a number of spectroscopic studies (UV/vis, NMR, EPR, resonance Raman, Mössbauer, X-ray absorption).^{11–16,77–80} Recently, even the elusive Cpd I of cytochrome P450 has been isolated in high yield and characterized spectroscopically as an iron(IV)-oxo porphyrin radical species.¹⁶ In view of the experimental consensus about the

iron(IV)-oxo porphyrin radical ground state for enzymatic Cpd I species and their synthetic analogues, the present prediction of low-lying iron(V)-oxo states from *ab initio* (RASPT2) calculations may be surprising and hard to believe. Even more, the calculations suggest an iron(V)-oxo ground state for both models, which seems to be contradicted by experimental results. However, when comparing the present calculations with experimental results, the following three factors should be taken into consideration.

First, even though the present RASPT2 calculations should be considered of high-quality—large basis sets and extensive active spaces were used, and the results obtained for the small model compare favorably to the CCSD(T) benchmark results—the calculated relative energies could still be in error by several kcal/mol. Uncertainties of such order may originate, for instance, from an unavoidable compromise between the size of the global active space and the size of the RAS2 subspace (for a discussion, see the Computational Details) as well as from the intrinsic approximations of RASPT2 theory. Because the various electromeric states are very close-lying, particularly for FeO(P)Cl, even a minor error in the calculated relative energies may easily change the character of the ground state to an iron(IV)-oxo porphyrin radical.

Second, the models studied here are clearly oversimplified and may not be directly comparable to experimentally investigated systems, the enzymes in particular. Moreover, the synthetic iron porphyrin complexes are based on porphyrin rings substituted at *meso* positions and some of them also at pyrrole β positions. The substituents, not included in the present models, may clearly affect the relative energy of iron(IV)-oxo porphyrin radicals and iron(V)-oxo states through their polar effect and/or by means of ruffling or saddling distortions of the porphyrin ring. The present choice of models was based on computational considerations, i.e., their high symmetry and relative simplicity, which are important factors in performing high-level *ab initio* calculations. This, in turn, is motivated by the main focus of the present study, which is benchmarking methods rather than making direct predictions about the situation in enzymes or for particular synthetic complexes.

Finally, the present calculations refer to isolated molecules *in vacuo*, while all known experiments probe the situation in solution or inside a protein pocket. On the basis of an implicit solvent model (COSMO), it was found here that the polarizable environment is an important factor favoring the iron(IV)-oxo porphyrin radical cation with respect to the iron(V)-oxo states. In fact, after considering the effect of a polar medium, the correct iron(IV)-oxo porphyrin radical ground state for FeO(P)Cl is recovered, in agreement with the experimental data for FeO-(TMP)Cl (where TMP = tetramesitylporphyrin).^{12,79,80} We further note that a qualitatively similar effect is known in P450 chemistry for iron species at one lower oxidation state (IV/III): the radical intermediate in the C–H hydroxylation mechanism has, *in vacuo*, a ground state with iron(IV) and a closed-shell porphyrin, (P)Fe^{IV}(OH)/Alk[•] (Alk[•] = alkyl radical). In contrast, the electromer with iron(III) and a porphyrin radical, (P^{•+})Fe^{III}(OH)/Alk[•], becomes most stable in a protein environment.¹

Interestingly, the RASPT2 calculations for FeO(P)⁺ point to an iron(V)-oxo ground state also in a polarizable medium. Even if, due to their limited precision (*vide supra*), these calculations may fail to predict the correct ground state of FeO(P)⁺, theory seems to suggest that the “naked”, five-coordinate cationic species, FeO(P)⁺, has a stronger preference for the iron(V)-oxo state than

the six-coordinated complex, FeO(P)Cl. In this regard, one may expect the actual coordination number of iron (five or six) to be an important experimental factor. Unfortunately, in many of the experimental studies dealing with complexes that are (formally) described as FeO(TMP)⁺,^{81–83} the actual ligation state of iron was not determined.¹² Since the chemical oxidants used to prepare the samples as well as the counterions from the iron(III) porphyrin substrates are potential axial ligands,¹² it is not clear whether the iron-oxo porphyrins probed in these experiments were isolated five-coordinate cations (thus comparable to the present five-coordinate model, FeO(P)⁺) or six-coordinate complexes (thus more comparable to the present six-coordinate model, FeO(P)Cl). One should also notice that some five-coordinated iron-oxo porphyrin species were produced in gas-phase experiments, and their reactivity was characterized.^{84,85} Although these species were tentatively assigned as iron(IV)-oxo porphyrin radical cations, their full spectroscopic characterization has not been performed, thus not permitting one to fully confirm this description.

The nature of the *ground state* has been quite firmly established for various Cpd I systems and their analogues. However, the character of the *low-lying excited states* and their potential role in multistate reactivity is less certain. The present theoretical study, likewise the previous one from the Jerusalem group,²⁴ clearly suggests that iron(IV)-oxo pentaradicaloid states and iron(V)-oxo states may be lying only a few kcal/mol above the triradicaloid ground state. Some recent experimental studies with the laser flash photolysis (LFP) technique¹⁹ certainly go in this direction (note, however, a recent criticism of the LFP experiments for P450 and CPO enzymes^{88–90}).

In particular, LFP experiments with metastable porphyrin-iron(IV) diperchlorates showed very reactive transients with oxo-transfer capabilities.^{20,21} Although the full spectroscopic characterization (e.g., EPR, Mössbauer) of these short-living intermediates was not possible, they were tentatively identified as iron(V)-oxo porphyrin species, on the basis of their unique UV/vis spectrum and very high reactivity in oxo-transfer reactions.²¹ Similar experimental data obtained from LFP experiments on iron corroles were also interpreted in terms of iron(V)-oxo corrole species.⁹¹ Newcomb et al. further argue that the putative iron(V)-oxo electromer might be initially produced by heterolytic cleavage of the FeO–ClO₃ bond, with a kinetic barrier preventing it from immediate conversion (by internal electron transfer) to a more stable iron(IV)-oxo porphyrin radical species.^{21,86} In this respect, they notice considerably high barriers for the analogous electronic isomerization of metastable Fe^{IV}O(P) to more stable Fe^{III}O(P^{•+}) species.⁹²

According to the present theoretical study, the iron(V)-oxo electromer is a good candidate for the low-lying excited form of high valent iron-oxo porphyrins that might be observed in these experiments. The alternative candidate suggested by theory is a pentaradicaloid iron(IV)-oxo species. However, the UV/vis spectra of the short-living intermediates lack similarity to known spectra of porphyrin radical cations,²¹ thus precluding this possibility and rather suggesting iron(V)-oxo character. The full interpretation of the LFP experiments^{20,21,91} and determination of a possible role for the putative iron(V)-oxo species in catalysis would require more elaborate spectroscopic studies and further theoretical calculations. Clearly, even if an excited iron(V)-oxo electromer could be generated photochemically, this does not automatically mean that this species should also be involved in actual oxygenation pathways.⁹³ However, this intriguing proposition

should not be excluded solely on the basis of the very high energy found for the iron(V)-oxo species in previous B3LYP calculations.^{2,23} The present study clearly indicates that this particular aspect of the energetics may not be well described by B3LYP and that the iron(V)-oxo states are expected at energies less than 10 kcal/mol above the ground state, even in a polar medium.

4. CONCLUSIONS

Two heme complexes with an iron-oxo group, FeO(P)⁺ and FeO(P)Cl (P = porphyrin), were studied in this work, employing a number of DFT methods as well as second-order perturbation theory based on either Complete Active Space (CASPT2) and Restricted Active Space (RASPT2) wave functions. The aim was to provide an accurate description of the energetics of various electromeric states of the studied systems and to access the accuracy of the computational methods. Further calibration of the predicted energetics was carried out with respect to coupled cluster CCSD(T) calculations for the small model system, FeO(L₂)⁺, containing two amidine ligands (η^2 -N₂C₃H₅⁻) instead of porphyrin.

The most important and intriguing suggestion from the present ab initio calculations undoubtedly concerns the presence of a low-lying iron(V)-oxo electromer. An iron(V)-oxo ground state is predicted by the present RASPT2 calculations for both heme models *in vacuo*. This is not consistent with the experimental observation of an iron(IV)-oxo porphyrin radical ground state in Cpd I species and their synthetic analogues.^{11–15,77,78} However, it was found here that the presence of a polarizable medium (a solvent or protein environment) stabilizes the states with iron(IV)-oxo porphyrin radical character (by up to 8 kcal/mol) with respect to the iron(V)-oxo states. Upon considering this effect, the iron(IV)-oxo porphyrin radical electromer is predicted for FeO(P)Cl in agreement with experimental data of similar complexes,^{12,79,80} but the iron(V)-oxo electromer is still low-lying and thermally accessible. A stronger preference for the iron(V)-oxo structure is predicted for the five-coordinate FeO(P)⁺ cation without a *trans* axial ligand. A comparison between the RASPT2 and CCSD(T) results for the small model indicates that both methods provide a consistent picture of the relative energetics for iron(V)-oxo and iron(IV)-oxo electromers. Therefore, no large, systematic errors are to be expected in the RASPT2 energies for the heme complexes. However, uncertainties of several kcal/mol are very difficult to avoid in computational ab initio studies on such complicated systems. Even such small errors in the calculated energies may easily change the identity of the predicted ground state. A second important conclusion is that the high-spin (pentaradicaloid) form of the iron(IV)-oxo porphyrin radical electromer is also low-lying, most probably about 10 kcal/mol or less above the triradicaloid iron(IV)-oxo state. This conclusion is valid even though we believe (on the basis of the comparison with the coupled cluster calculations and on previous experience) that the CASPT2 and RASPT2 calculations may somewhat overestimate the stability of the high-spin state.

The role of the low-lying pentaradicaloid iron(IV)-oxo and the putative iron(V)-oxo electromer for mechanisms of catalytic reactions is currently under debate.^{2,21,24,88,93} Computational studies have already shown that the pentaradicaloid states have a lower barrier for H-abstraction than the triradicaloids.^{23,24,65} On the basis of experimental suggestions,²¹ a similar enhanced reactivity was intuitively proposed for the excited iron(V)-oxo electromer, although such a proposal has so far not been

definitely confirmed or rejected. The true role of the excited iron(V)-oxo electromers for catalytic properties of high-valent iron-oxo porphyrins and corroles remains to be addressed by further experimental and theoretical work. For further theoretical studies with DFT, it might be important to remember the dramatic difference between the hybrid and nonhybrid functionals in regard to (a) the spin promotion energy on the iron (i.e., difference in energy between tri- and pentaradicaloid iron(IV)-oxo states) and (b) the relative energy of the iron(V)-oxo vs iron(IV)-oxo porphyrin radical states. While comparison with ab initio calculations—in particular, with the CCSD(T) benchmark results for the small model—essentially confirms the high accuracy of hybrid functionals in predicting the spin promotion energy, these functionals seem to be much less accurate in describing the relative position of the iron(V)-oxo states. In fact, for the presently studied systems, it was quite difficult to find a single DFT method that describes both aspects of the energetics reasonably well (assuming, of course, that the present ab initio results are treated as a benchmark); among the functionals tested here, only OLYP points to relative energetics in accord with the ab initio picture.

We believe that the present study has also demonstrated some advantages of RASPT2 over standard CASPT2 calculations. The RASPT2 calculations performed in this study were based on a fairly large active space including not only all important orbitals on the FeO group but also a large number of porphyrin π orbitals and their correlating π^* orbitals. This is in contrast to the CASPT2 calculations where no more than just a few frontier orbitals of the porphyrin may be included in the active space (such like the Gouterman set of two π and two π^* orbitals). It turned out that the CASPT2 calculations cannot reliably reproduce (a) the splitting between a_{2u} and a_{1u} radicals on the porphyrin or (b) the magnetic coupling in the triradicaloid iron(IV)-oxo states. Both problems are readily solved in the present RASPT2 calculations with the larger active space. Furthermore, due to the better description of electron transfer from the iron to the porphyrin fragment, the relative energy of the iron(V)-oxo and iron(IV)-oxo ligand radical electromers is expected to be improved in the RASPT2 calculations. The RASPT2 method thus appears to be a promising tool for studying difficult cases with transition metals where very large active spaces may be unavoidable.

■ ASSOCIATED CONTENT

S Supporting Information. Optimized Cartesian coordinates, additional DFT results, contour plots of the 29in28 active orbitals, more technical details and the test results obtained from CASSCF/CASPT2 and RASSCF/RASPT2 calculations. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: mradon@chemia.uj.edu.pl.

■ ACKNOWLEDGMENT

This research project has been supported by grants from the Flemish Science Foundation (FWO), the Concerted Research Action of the Flemish Government (GOA), and the funds for

scientific research of the Ministry of Science and Higher Education of Poland (MNiSW; grant number N N204 333837). The computational grants from the Academic Computer Center in Kraków (CYFRONET) and in Gdańsk (CI TASK) are also acknowledged. Molecular graphics in Figure 2 were prepared with the XYZViewer program obtained from Sven de Marothy (Stockholm University).

REFERENCES

- (1) Shaik, S.; Kumar, D.; de Visser, S. P.; Altun, A.; Thiel, W. *Chem. Rev.* **2005**, *105*, 2279–2328.
- (2) Shaik, S.; Cohen, S.; Wang, Y.; Chen, H.; Kumar, D.; Thiel, W. *Chem. Rev.* **2010**, *110*, 949–1017.
- (3) Pan, Z.; Zhang, R.; Newcomb, M. *J. Inorg. Biochem.* **2006**, *100*, 524–532.
- (4) de Oliveira, F. T.; Chanda, A.; Banerjee, D.; Shan, X.; Mondal, S.; Lawrence Que, J.; Bominaa, E. L.; Münck, E.; Collins, T. *J. Science* **2007**, *315*, 835–838.
- (5) Lee, S.; Han, J.; Kwak, H.; Lee, S.; Lee, E.; Kim, H.; Lee, J.; Bae, C.; Lee, S.; Kim, Y.; Kim, C. *Chem.—Eur. J.* **2007**, *13*, 9393–9398.
- (6) Wasbotten, I.; Ghosh, A. *Inorg. Chem.* **2006**, *45*, 4910–4913.
- (7) Dey, A.; Ghosh, A. *J. Am. Chem. Soc.* **2002**, *124*, 3206–3207.
- (8) Ghosh, A. *J. Biol. Inorg. Chem.* **2006**, *11*, 712–724.
- (9) It should be further noted that in the case of the P450 Cpd I, the ligand radical resides in a combination of a_{2u} with the σ orbital of sulphur from the axial cysteine. See, e.g., ref 1.
- (10) Obviously, the term “porphyrin radical cation” and the notation “ $P^{•+}$ ” refer both to an open-shell porphyrin species with a charge of -1 , as compared to a closed-shell porphyrin species with a charge of -2 (formally a porphyrinate anion). This notation is customary for high-valent metal-oxo species with noninnocent macrocyclic ligands—see for instance refs 2, 19–22, 24.
- (11) Gold, A.; Weiss, R. *J. Porph. Phthal.* **2000**, *4*, 344–349.
- (12) Weiss, R.; Bulach, V.; Gold, A.; Terner, J.; Trautwein, A. *J. Biol. Inorg. Chem.* **2001**, *6*, 831–845.
- (13) Groves, J. T. *J. Inorg. Biochem.* **2006**, *100*, 434–447.
- (14) Kim, S. H.; Perera, R.; Hager, L. P.; Dawson, J. H.; Hoffman, B. M. *J. Am. Chem. Soc.* **2006**, *128*, 5598–5599.
- (15) Nam, W. *Acc. Chem. Res.* **2007**, *40*, 522–531.
- (16) Rittle, J.; Green, M. T. *Science* **2010**, *330*, 933–937.
- (17) Yamaguchi, K.; Watanabe, Y.; Morishima, I. *J. Chem. Soc., Chem. Commun.* **1992**, 1721–1723.
- (18) Murakami, T.; Yamaguchi, K.; Watanabe, Y.; Morishima, I. *Bull. Chem. Soc. Jpn.* **1998**, *71*, 1343–1353.
- (19) Zhang, R.; Newcomb, M. *Acc. Chem. Res.* **2008**, *41*, 468–477.
- (20) Pan, Z.; Zhang, R.; Fung, L. W.-M.; Newcomb, M. *Inorg. Chem.* **2007**, *46*, 1517–1519.
- (21) Pan, Z.; Wang, Q.; Sheng, X.; Horner, J. H.; Newcomb, M. *J. Am. Chem. Soc.* **2009**, *131*, 2621–2628.
- (22) Ogliaro, F.; de Visser, S. P.; Groves, J. T.; Shaik, S. *Angew. Chem., Int. Ed.* **2001**, *40*, 2874–2878.
- (23) Altun, A.; Shaik, S.; Thiel, W. *J. Am. Chem. Soc.* **2007**, *129*, 8978–8987.
- (24) Chen, H.; Song, J.; Lai, W.; Wu, W.; Shaik, S. *J. Chem. Theory Comput.* **2010**, *6*, 940–953.
- (25) Altun, A.; Kumar, D.; Neese, F.; Thiel, W. *J. Phys. Chem. A* **2008**, *112*, 12904–12910.
- (26) Radoń, M.; Broclawik, E. *J. Chem. Theory Comput.* **2007**, *3*, 728–734.
- (27) Schöneboom, J. C.; Neese, F.; Thiel, W. *J. Am. Chem. Soc.* **2005**, *127*, 5840–5853.
- (28) Andersson, K.; Malmqvist, P.-Å.; Roos, B. O. *J. Chem. Phys.* **1991**, *96*, 1218–1226.
- (29) Malmqvist, P.-Å.; Pierloot, K.; Shahi, A. R. M.; Cramer, C. J.; Gagliardi, L. *J. Chem. Phys.* **2008**, *128*, 204109.
- (30) Pierloot, K.; Zhao, H.; Vancoillie, S. *Inorg. Chem.* **2010**, *49*, 10316–10329.
- (31) Strickland, N.; Harvey, J. N. *J. Phys. Chem. B* **2007**, *111*, 841–852.
- (32) Olah, J.; Harvey, J. *J. Phys. Chem. A* **2009**, *113*, 7338–7345.
- (33) Ahlrichs, R.; Horn, H.; Schaefer, A.; Treutler, O.; Haeser, M.; Baer, M.; Boecker, S.; Deglmann, P.; Furche, F. *Turbomole v5.9*; Universitaet Karlsruhe; Karlsruhe, Germany.
- (34) 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.
- (35) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.
- (36) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (37) Perdew, J. P.; Ernzerhof, M.; Burke, K. *J. Chem. Phys.* **1996**, *105*, 9982–9985.
- (38) Reiher, M.; Salomon, O.; Hess, B. A. *Theor. Chem. Acc.* **2001**, *107*, 48–55.
- (39) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.
- (40) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822–8824.
- (41) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- (42) Handy, N. C.; Cohen, A. *J. Mol. Phys.* **2001**, *99*, 403–412.
- (43) Tawada, Y.; Tsuneda, T.; Yanagisawa, S.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2004**, *120*, 8425.
- (44) Vydrov, O. A.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 234109.
- (45) Vydrov, O. A.; Heyd, J.; Krukau, A. V.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 074106.
- (46) Vydrov, O. A.; Scuseria, G. E.; Perdew, J. P. *J. Chem. Phys.* **2007**, *126*, 154109.
- (47) Klamt, A. *J. Phys. Chem.* **1995**, *99*, 2224–2235.
- (48) Cossi, M.; Rega, N.; Scalmani, G.; Barone, V. *J. Chem. Phys.* **2001**, *114*, 5691–5701.
- (49) Karlström, G.; Lindh, R.; Malmqvist, P.-Å.; Roos, B.; Ryde, U.; Veryazov, V.; Widmark, P.-O.; Cossi, M.; Schimmelpfennig, B.; Neogrady, P.; Sejro, L. *Comput. Mater. Sci.* **2003**, *28*, 222–239; see: <http://www.teokem.lu.se/molcas> (accessed March 2011).
- (50) Reiher, M.; Wolf, A. *J. Chem. Phys.* **2004**, *121*, 10945–10956.
- (51) Ghigo, G.; Roos, B.; Malmqvist, P.-Å. *Chem. Phys. Lett.* **2004**, *396*, 142–149.
- (52) Aquilante, F.; Malmqvist, P.-Å.; Pedersen, T. B.; Ghosh, A.; Roos, B. O. *J. Chem. Theory Comput.* **2008**, *4*, 694–702.
- (53) Roos, B. O.; Lindh, R.; Malmqvist, P.-Å.; Veryazov, V.; Widmark, P.-O. *J. Phys. Chem. A* **2005**, *109*, 6575–6579.
- (54) Pierloot, K.; Dumez, B.; Widmark, P.-O.; Roos, B. *Theor. Chim. Acta* **1995**, *90*, 87–114.
- (55) Roos, B. O.; Andersson, K.; Fulscher, M.; Malmqvist, P.-Å.; Serrano-Andres, L.; Pierloot, K.; Merchán, M. Multiconfigurational perturbation theory: Applications in 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. 93.
- (56) Pierloot, K. Nondynamic Correlation Effects in Transition Metal Coordination Compounds. In *Computational Organometallic Chemistry*; Cundari, T. R., Ed.; Marcel Dekker, Inc.: New York, 2001.
- (57) Pierloot, K. *Mol. Phys.* **2003**, *101*, 2083–2094.
- (58) Sauri, V.; Serrano-Andres, L.; Shahi, A. R. M.; Gagliardi, L.; Vancoillie, S.; Pierloot, K. *J. Chem. Theory Comput.* **2011**, *7*, 153–168.

(59) 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.; Püger, K.; Pitzer, R.; Reiher, M.; Shiozaki, T.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T.; Wang, M.; Wolf, A. *MOLPRO*, version 2009. See <http://www.molpro.net> (accessed March 2011).

(60) de Jong, W. A.; Harrison, R. J.; Dixon, D. A. *J. Chem. Phys.* **2001**, *114*, 48–53.

(61) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007–1023.

(62) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *J. Chem. Phys.* **1997**, *106*, 9639–9646.

(63) Harvey, J. N.; Aschi, M. *Faraday Discuss.* **2003**, *124*, 129–143.

(64) Beran, G. J. O.; Gwaltney, S. R.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2003**, *5*, 2488–2493.

(65) Chen, H.; Lai, W.; Shaik, S. *J. Phys. Chem. Lett.* **2010**, *1*, 1533–1540.

(66) Lee, T. J.; Taylor, P. R. *Int. J. Quantum Chem.* **1989**, *36*, 199–207.

(67) Janssen, C. L.; Nielsen, I. M. B. *Chem. Phys. Lett.* **1998**, *290*, 423–430.

(68) Ogliaro, F.; Cohen, S.; de Viser, S. P.; Shaik, S. *J. Am. Chem. Soc.* **2000**, *122*, 12892–12893.

(69) Harvey, J. N. *Annu. Rep. Prog. Chem., Sect. C: Phys. Chem.* **2006**, *102*, 203–226.

(70) Pierloot, K.; Vancoillie, S. *J. Chem. Phys.* **2006**, *125*, 124303.

(71) Pierloot, K.; Vancoillie, S. *J. Chem. Phys.* **2008**, *128*, 034104.

(72) Kepenekian, M.; Robert, V.; Le Guennic, B. *J. Chem. Phys.* **2009**, *131*, 114702.

(73) Radoń, M.; Broclawik, E.; Pierloot, K. *J. Phys. Chem. B* **2010**, *114*, 1518–1528.

(74) Vancoillie, S.; Zhao, H.; Radoń, M.; Pierloot, K. *J. Chem. Theory Comput.* **2010**, *6*, 576–582.

(75) Neogrády, P.; Medved', M.; Černušák, I.; Urban, M. *Mol. Phys.* **2002**, *100*, 541–560.

(76) Karton, A.; Rabinovich, E.; Martin, J. M. L.; Ruscic, B. *J. Chem. Phys.* **2006**, *125*, 144108.

(77) Groves, J. T.; Haushalter, R. C.; Nakamura, M.; Nemo, T. E.; Evans, B. J. *J. Am. Chem. Soc.* **1981**, *103*, 2884–2886.

(78) Hosten, C. M.; Sullivan, A. M.; Palaniappan, V.; Fitzgerald, M. M.; Terner, J. *J. Biol. Chem.* **1994**, *269*, 13966–13978.

(79) Gross, Z.; Nimri, S. *J. Am. Chem. Soc.* **1995**, *117*, 8021–8022.

(80) Wolter, T.; Meyer-Klaucke, W.; Müther, M.; Mandon, D.; Winkler, H.; Trautwein, A. X.; Weiss, R. *J. Inorg. Biochem.* **2000**, *78*, 117–122.

(81) Mandon, D.; Weiss, R.; Jayaraj, K.; Gold, A.; Bill, E.; Trautwein, A. X. *Inorg. Chem.* **1992**, *31*, 4404–4409.

(82) Müther, M.; Bill, E.; Trautwein, A. X.; Mandon, D.; Weiss, R.; Gold, A.; Jayaraj, K.; Austin, R. N. *Hyperfine Interact.* **1994**, *91*, 803–808.

(83) Jones, R.; Jayaraj, K.; Gold, A.; Kirk, M. L. *Inorg. Chem.* **1998**, *37*, 2842–2843.

(84) Crestoni, M. E.; Fornarini, S. *Inorg. Chem.* **2007**, *46*, 9018–9020.

(85) Chiavarino, B.; Cipollini, R.; Crestoni, M.; Fornarini, S. *J. Am. Chem. Soc.* **2008**, *130*, 3208–3217.

(86) Sheng, X.; Horner, J. H.; Newcomb, M. *J. Am. Chem. Soc.* **2008**, *130*, 13310–13320.

(87) Wang, Q.; Sheng, X.; Horner, J. H.; Newcomb, M. *J. Am. Chem. Soc.* **2009**, *131*, 10629–10636.

(88) Rittle, J.; Younker, J. M.; Green, M. T. *Inorg. Chem.* **2010**, *49*, 3610–3617.

(89) Yuan, X.; Sheng, X.; Horner, J. H.; Bennett, B.; Fung, L. W.-M.; Newcomb, M. *J. Inorg. Biochem.* **2010**, *104*, 1156–1163.

(90) It must be noted here that the LFP experiments for cytochrome P450 and CPO enzyme (refs 86, 87) have been recently criticized (refs 88, 89) as not generating Cpd I despite the initial claims of the authors.

In these studies, an enzyme was treated first with peroxynitrite to generate Cpd II, and subsequent photooxidation was (incorrectly) expected to yield Cpd I. However, in this study, we mostly refer to other LFP experiments to which the recent criticism does not seem to apply—namely, the study described in refs 20 and 21. This study was carried out not for an enzyme but for a small biomimetic complex, and it employed a different oxidant agent, namely, iron(III) perchlorate. Moreover, the intermediate species was not a Cpd II analogue but an iron(IV) porphyrin diperchlorate, whose subsequent LFP gave a short-living transient that was tentatively assigned as iron(V)-oxo porphyrin perchlorate (cf. refs 20, 21). We believe that the criticism of some LFP experiments expressed by Green et al. (ref 88) remains specific to those experiments with enzymes, while LFP in general remains a well-established technique for generating high-valent metal-oxo complexes and for studying their reactivity (cf. ref 19).

(91) Harischandra, D.; Zhang, R.; Newcomb, M. *J. Am. Chem. Soc.* **2005**, *127*, 13776–13777.

(92) Pan, Z.; Harischandra, D. N.; Newcomb, M. *J. Inorg. Biochem.* **2009**, *103*, 174–181.

(93) de Montellano, P. R. O. *Chem. Rev.* **2010**, *110*, 932–948.

Projected Coupled Cluster Amplitudes from a Different Basis Set As Initial Guess

Marco Caricato,* Gary W. Trucks, and Michael J. Frisch

Gaussian, Inc., 340 Quinipiac St. Bldg. 40, Wallingford, Connecticut 06492, United States

S Supporting Information

ABSTRACT: Many model chemistry schemes require a sequence of high level calculations, often at the coupled cluster (CC) level, with increasingly larger basis sets. The CC equations are solved iteratively, and the rate of convergence strongly depends on the quality of the initial guess. Here, we propose a strategy to define a better guess from a previous CC calculation with a different basis set by employing the concept of corresponding orbitals.^{1,2} Only the part of the converged amplitudes from the previous calculation that has a large correspondence to the space spanned by the new basis set is projected and used as a new starting point. The computational time for the projection is negligible and significantly reduces the number of cycles necessary for convergence in comparison to the standard guess based on the first order wave function. Numerical results are presented for ground and excited state calculations with the CC singles and doubles (CCSD) and the equation of motion CCSD (EOM-CCSD) methods with the restricted and unrestricted Hartree–Fock (HF) reference functions. However, this approach is more general and can be applied to any truncation of the cluster expansion and reference function.

1. INTRODUCTION

Single reference coupled cluster (CC) methods currently represent the most computationally balanced level of theory in quantum chemistry to obtain accurate and consistent descriptions of ground and excited state wave functions for small- and medium-sized molecules.³ Various levels of truncation of the wave operator can be defined and used, often depending on what is computationally affordable for a particular system. Among these, the most widely used approximation of the complete excitation manifold includes single and double excitation operators (CCSD).⁴ Another factor that plays an important role in the accuracy of the calculation is the finite one-electron expansion basis set. Indeed, CC methods show a significant sensitivity to the quality of the basis set, and often large bases are necessary to reach convergence in the description of properties.⁵ This is particularly true for excited states where diffuse functions are frequently unavoidable even for a qualitatively sound description of electronic transitions. Also in this respect, performing production calculations in a reasonable time poses a limit to the size of the basis sets that can be effectively used.

In recent years, many efforts have been carried out to extend the range of applications of CC theory to larger systems and/or larger basis expansions.³ Examples of these include introducing further approximations in the wave operator,^{6–8} combining various levels of theory in hybrid methods,^{9–11} considering explicit electron correlation to improve basis set convergence,^{12–14} and defining protocols for thermochemistry and basis set extrapolations.^{15–20} In most cases, CC equations are solved iteratively, stopping when certain convergence criteria are met. The rate of convergence is fundamentally influenced by two factors: the quality of the initial guess and the extrapolation technique used to define the amplitudes in the next iterative cycle. The latter is normally dealt with using standard techniques

like the direct inversion of the iterative subspace (DIIS)^{21–23} or the reduced linear equation (RLE)^{24,25} methods for ground state calculations and extended Davidson algorithms for excited states.^{26–28} In this article, we concentrate on the former issue, i.e., how to define a better starting point for a CC calculation assuming that a former calculation with another, usually smaller, basis set has already been performed. Although we will focus the discussion on ground state CCSD and equation of motion CCSD (EOM-CCSD)²⁹ for excited states with the canonical form of the restricted and unrestricted Hartree–Fock (RHF/UHF) wave functions as a reference, our approach is general and can be applied to other truncated CC wave functions.

This approach is useful when multiple calculations with various basis sets are required, such as basis set extrapolation methods and model chemistry protocols.^{15–19} Another example is excited state studies when exploratory calculations on many states are performed with a smaller basis that can provide semiquantitative results, and then the calculation for some specific state is refined with a larger basis set. We provide examples of these type of applications in this paper. Although no example is reported here, the application of such methodology to the Brueckner doubles (BD)^{30,31} variant of CCSD (or to geometry optimizations) provides large savings in computational time as the orbitals are rotated at each macroiteration step. It can even be used in the reverse order, i.e., passing from a larger to a smaller basis, for instance, for investigating higher excited states with a less demanding basis set after converging lower ones.

This paper is organized as follows: The theory is presented in section 2, the results of the test calculations are reported in section 3, and concluding remarks are discussed in section 4.

Received: November 7, 2010

Published: February 28, 2011

2. THEORY

The usual guess used for ground state CCSD is to set to zero the singles amplitudes, t_i^a (when the canonical orbitals are used and the Fock matrix is diagonal), and use the first order wave function for the doubles amplitudes:

$$t_{ab}^{ij(1)} = \frac{\langle ij||ab \rangle}{\Delta_{ij}^{ab}} \quad (1)$$

where i, j and a, b represent occupied and virtual orbitals, respectively, $\langle ij||ab \rangle$ are the antisymmetrized two-electron integrals, and $\Delta_{ij}^{ab} = f_{ii} + f_{jj} - f_{aa} - f_{bb}$ is the combination of the diagonal elements of the Fock matrix (orbital energies for canonical orbitals). This is a sensible way to define the guess, but it may require many cycles to reach convergence if the second order energy from the perturbation theory expansion is not a good approximation of the correlation energy. For the evaluation of molecular properties, the calculation of the electronic density is required. In the case of CC wave functions, another set of amplitudes is necessary, called Λ - or Z -vector amplitudes,^{32,33} that can use the converged t amplitudes as an initial guess. For EOM-CCSD, two sets of amplitudes (for each excited state) are necessary for the calculation of transition energies and properties that correspond to the left and right eigenvectors of the EOM-CCSD Hamiltonian. It has been shown that the most efficient way to compute them is to solve for the two sets separately.^{26–28} For the solution of the first set, usually the right eigenvectors, the converged configuration interaction singles (CIS) eigenvectors are used for the guess. Since the CIS eigenvectors only include single excitations, the double excitation component of the EOM-CCSD eigenvectors are initialized to zero. When the right eigenvectors are found, they are used as the guess for the left ones. Although the CIS vectors are a sensible choice given that EOM-CCSD mainly provides an accurate description of one-electron transitions where the contribution of single excitations is predominant, the convergence can be slow if double excitation contributions are significant.

We propose to project the converged amplitudes from a previous calculation with a certain basis set onto the space spanned by a new basis set and use the projected amplitudes as a starting guess. This requires determining which part of the space in the new basis expansion is spanned by the old basis. This can be achieved using *corresponding orbitals*.^{1,2} In the following, we assume real orbitals. We first define the overlap matrices \mathbf{S} between the occupied molecular orbitals (MOs) and $\bar{\mathbf{S}}$ between the virtual MOs in the two basis sets:

$$\begin{aligned} S_{ij} &= \langle \tilde{i}|j \rangle = \sum_{\mu, \nu} c_{\mu i} \tilde{c}_{\mu j} \langle \tilde{\mu}|\nu \rangle c_{\nu j} \\ \bar{S}_{ab} &= \langle \tilde{a}|b \rangle = \sum_{\mu, \nu} c_{\mu a} \tilde{c}_{\mu b} \langle \tilde{\mu}|\nu \rangle c_{\nu b} \end{aligned} \quad (2)$$

where the indexes with the symbol \sim refer to the *old* basis set, c represents the MO coefficients, and μ and ν are the atomic orbitals (AOs). We assume the MOs to be orthonormalized within each set. The matrices \mathbf{S} and $\bar{\mathbf{S}}$ are then decomposed using the single value decomposition (SVD) technique:³⁴

$$\begin{aligned} \mathbf{S} &= \mathbf{U}\sigma\mathbf{V}^\dagger \\ \bar{\mathbf{S}} &= \bar{\mathbf{U}}\bar{\sigma}\bar{\mathbf{V}}^\dagger \end{aligned} \quad (3)$$

where \mathbf{U} and \mathbf{V}^\dagger and $\bar{\mathbf{U}}$ and $\bar{\mathbf{V}}^\dagger$ are unitary transformations of the *old* and *new* occupied and virtual MOs, and σ and $\bar{\sigma}$ are diagonal matrices. This does not affect the wave function, as it is invariant

under unitary transformations of the occupied–occupied and virtual–virtual blocks of MOs. The values of σ and $\bar{\sigma}$ are $0 \leq \sigma_{ii}\bar{\sigma}_{aa} \leq 1$ (the number of nonzero values is equal to the size of the smaller of the MO sets). The transformations \mathbf{U} , \mathbf{V}^\dagger , $\bar{\mathbf{U}}$, and $\bar{\mathbf{V}}^\dagger$ can be interpreted as the rotations of the two basis sets into two sets that most correspond to each other. If the values of σ and $\bar{\sigma}$ are organized in descending order, we can define the projection matrices \mathbf{P} and $\bar{\mathbf{P}}$ and the extraction matrices \mathbf{E} and $\bar{\mathbf{E}}$ as

$$\begin{aligned} P_{ij} &= \sum_k^{\text{corr}} U_{ik} V_{jk} \\ \bar{P}_{ab} &= \sum_c^{\text{corr}} \bar{U}_{ac} \bar{V}_{bc} \end{aligned} \quad (4)$$

$$\begin{aligned} E_{ij} &= \sum_k^{\text{corr}} V_{ik} V_{jk} \\ \bar{E}_{ab} &= \sum_c^{\text{corr}} \bar{V}_{ac} \bar{V}_{bc} \end{aligned} \quad (5)$$

where the indexes k and c run over the *corresponding orbitals* with σ_{kk} and $\bar{\sigma}_{cc} \geq 0.9$. The projection matrices are used to map the amplitudes in the old basis set onto the new basis. The purpose of the matrices \mathbf{E} and $\bar{\mathbf{E}}$, when applied onto an orbital i or a , is to extract the contribution of such an orbital to the selected *corresponding orbitals*. Two sets of projection and extraction matrices are defined separately for electrons with spin α and spin β for an unrestricted reference function. The new guess amplitudes can be calculated as

$$t_i^a = \sum_{\bar{i}\bar{a}} \bar{P}_{\bar{i}\bar{a}} t_{\bar{i}}^{\bar{a}} P_{\bar{i}\bar{a}} \quad (6)$$

$$t_{ij}^{ab} = t_{ij}^{ab(1)} - \sum_{klcd} \bar{E}_{ca} \bar{E}_{db} t_{kl}^{cd(1)} E_{ki} E_{lj} + \sum_{\bar{i}\bar{j}\bar{a}\bar{b}} \bar{P}_{\bar{i}\bar{a}} \bar{P}_{\bar{j}\bar{b}} t_{\bar{i}\bar{j}}^{\bar{a}\bar{b}} P_{\bar{i}\bar{j}} \quad (7)$$

where $t_{ij}^{ab(1)}$ is the first order wave function as in eq 1. When the new guess for the amplitudes t_{ij}^{ab} in eq 7 is built, only the part of the first order amplitudes in eq 1 that gives no contribution to the *corresponding orbitals* in the old basis set is retained, by applying \mathbf{E} and $\bar{\mathbf{E}}$. The contraction in eq 7 scales as $O(N^5)$, where N is the number of basis functions, since the indexes are contracted one at a time. The same transformations in eqs 6 and 7 are employed for the \mathbf{Z} vector and the EOM-CCSD eigenvector amplitudes. We used the expression in eq 1 to define the first order term for the projected \mathbf{Z} vector double amplitudes, whereas this is set to zero for EOM-CCSD. This procedure has been implemented in the Gaussian 09 suite of programs.³⁵

3. RESULTS

All of the data in this section is reported as the ratio of the number of iterative steps with the projected guess compared to the calculations with the normal guess. The absolute number of steps is reported in the Supporting Information. The extrapolation technique used to solve the ground state t and \mathbf{Z} vector equations is the RLE scheme of Purvis and Bartlett,²⁴ while an extended Davidson algorithm is used for the EOM-CCSD equations.^{27,28} The details of these algorithms are described in the literature and need not be repeated here. The choice of these algorithms is due to their efficiency for the solution of the CC equations compared to other options, as demonstrated in refs 24, 25, 27, and 28.

We start with two examples where this methodology is applied in production-like calculations using compound energy models. Section 3.1 reports the application of the W1U theory³⁶ on the phenyl and phenol radicals with and without the CC amplitude projection. The projected guess is also tested in ground state calculations in section 3.2. The test molecules include formaldehyde, butadiene, ethene, and acetone for the RHF reference and the vinyl radical for the UHF reference. The geometries are taken from ref 37 for the closed shell species and from ref 38 for the radical. The same systems and basis sets are used for the excited states in section 3.3. We consider one and three states per irrep (thus, a total of 4 and 12 states for formaldehyde, butadiene, and acetone; 2 and 6 for the vinyl radical; and 8 and 24 for ethene). All of these molecules have been thoroughly studied experimentally and theoretically.^{5,37–39} The values for the transition energies and oscillator strengths for all of the basis sets considered in this work can be found in the Supporting Information.

The convergence criteria for ground and excited state calculations are based on changes in both energy and wave function (the same criteria are used for all the basis sets). In particular, for the W1U case, convergence is achieved for changes $<10^{-7}$ au in the energy and $<10^{-5}$ in the wave function (t amplitudes). For the other ground state cases, the thresholds are 10^{-8} au for the energy and 10^{-6} for the wave function (t and Z amplitudes). For the EOM-CCSD calculations, the thresholds are 10^{-7} au for the excitation energy and 10^{-5} for the wave function (right and left eigenvectors).

Various sequences of basis sets are examined where each calculation uses the projected guess from the previous basis. This

Table 1. W1U Ratio of the Number of Cycles (For the Last Four Steps) and of the Total CPU Time with the Projected Amplitude Guess with Respect to the Standard Guess

	$C_6H_5^*$	$C_6H_5O^*$
step 4	0.68	0.70
step 5	0.63	0.65
step 6	0.63	0.78
step 7	0.68	0.67
total time	0.79	0.74

Table 2. Ground State, First Sequence of Basis Sets^a

	6-31+G ^b	6-31++G ^{**}	aug-cc-pVDZ	aug-cc-pVTZ	aug-cc-pVQZ
t amplitudes					
formaldehyde	0.87	0.80	0.87	0.80	0.73
ethene	0.85	0.85	0.85	0.85	0.77
butadiene	0.76	0.76	0.82	0.82	
acetone	0.94	0.76	0.82	0.82	
vinyl radical	0.84	0.78	0.89	0.75	
Z vector					
formaldehyde	1.08	0.85	0.92	0.92	0.85
ethene	0.85	0.92	0.85	0.85	0.71
butadiene	0.76	0.81	1.08	1.08	
acetone	0.94	0.76	0.82	0.93	
vinyl radical	1.07	1.00	0.94	0.82	

^aRatio of the number of cycles with the projected amplitude guess with respect to the standard guess. ^bInitial basis set: 6-31G*. Amplitudes are projected from the previous basis in the sequence.

sequences could be used, for instance, in basis set extrapolation techniques. We do not consider all of the possible combinations of basis sets as guesses for the others, as this would make the interpretation of the results confusing. The goal for these tests is to evaluate when a basis set may be a good starting point for projecting the amplitudes onto a larger basis. Therefore, we also include choices that may be not recommendable in production calculations (i.e., basis sets without diffuse functions for excited states). In these test cases, all the excited states are considered for all the basis sets, although it is more likely that in production calculations only a few states may be refined with a larger basis set after some exploratory calculations with a smaller basis.

3.1. W1U Tests. W1U³⁶ is the variation of the Weizmann-1 theory (W1)^{18,19} that uses the unrestricted CCSD(T)⁴⁰ method as the highest level of theory and is used to obtain highly accurate thermochemical data for open shell systems (average accuracy within 1 kcal/mol). The method consists of seven subcalculations:

- Step 1: Geometry optimization at the B3LYP level with the cc-pVTZ+d basis set
- Step 2: Frequency at the B3LYP level with the cc-pVTZ+d basis set
- Step 3: Energy calculation at the CCSD(T) level with the augh-cc-pVDZ+2df basis set
- Step 4: Energy calculation at the CCSD(T) level with the augh-cc-pVTZ+2df basis set
- Step 5: Energy calculation at the CCSD level with the augh-cc-pVQZ+2df basis set

Table 3. Ground State, Second Sequence of Basis Sets^a

	t amplitudes		Z vector	
	aug-cc-VDZ ^b	d-aug-cc-pVDZ	aug-cc-VDZ ^b	d-aug-cc-pVDZ
formaldehyde	0.93	0.73	1.00	0.85
ethene	0.85	0.69	0.85	0.90
butadiene	0.76	0.50	1.00	0.77
acetone	0.88	0.71	0.88	0.75
vinyl radical	0.89	0.63	0.94	0.80

^aRatio of the number of cycles with the projected amplitude guess with respect to the standard guess. ^bInitial basis set: cc-pVDZ. Amplitudes are projected from the previous basis in the sequence.

- Step 6: Energy calculation at the CCSD(T) level with the MTSmall basis set
- Step 7: Energy calculation at the CCSD(T) level with the

MTSmall basis set with correlation of the core electrons followed by an energy extrapolation. MTSmall consists of a completely decontracted cc-pVTZ basis set with two tight d and one tight f function added. We tested two radicals, $C_6H_5^\bullet$ and $C_6H_5O^\bullet$, with and without the projection scheme presented in section 2. The geometries and the energy information are reported in the Supporting Information as well as the total number of iterations for the CC subcalculations. The ratios for

Table 4. Ground State, Third Sequence of Basis Sets^a

	<i>t</i> amplitudes		Z vector	
	aug-cc-VTZ ^b	d-aug-cc-pVTZ	aug-cc-VTZ ^b	d-aug-cc-pVTZ
formaldehyde	0.87	0.67	0.92	0.71
ethene	0.85	0.69	0.85	0.82
vinyl radical	0.90	0.60	1.00	0.73

^aRatio of the number of cycles with the projected amplitude guess with respect to the standard guess. ^bInitial basis set: cc-pVTZ. Amplitudes are projected from the previous basis in the sequence.

the iterations for the steps 4–7 and for the total time are reported in Table 1. The savings in total time is smaller than for the iterative solution of the CCSD equations in the individual steps. This is because the perturbative triples correction to the CCSD energy in steps 4, 6, and 7 is not iterative, and therefore it does not benefit from the improved guess. Nonetheless, the savings in total CPU time is 20–25%, which is remarkable considering the length of this type of calculation. Hence, these test cases show a considerable reduction in the number of iterations and in the total calculation time when the projection scheme is used to define the guess for the next subcalculation. In fact, for open shell species with large spin contamination in the reference function, CCSD requires many iterations to remove the contribution of the first spin contaminant from the energy.^{41,42} Thus, starting with projected *t* amplitudes from another basis significantly reduces the number of iterations.

3.2. Ground State. The results for the ground state calculations are reported in Tables 2–4. The general trends show a significant reduction of the computational cost for the *t* amplitudes, in the range of 10–30%, with peaks of 40–50% in Tables 3 and 4. The savings is smaller for the Z vector, which indicates that the converged *t* amplitudes are a good starting point. The results

Table 5. Excited State, First Sequence of Basis Sets^a

	6-31+G ^{*b}		Right Eigenvector		
	6-31++G ^{**}	aug-cc-pVDZ	aug-cc-pVTZ	aug-cc-pVQZ	
			three states/irrep		
formaldehyde	0.97	0.90	0.89	0.91	0.86
ethene	0.96	0.88	0.93	0.92	0.86
butadiene	1.00	0.86	0.87	0.87	
acetone	1.01	0.91	0.86	0.85	
vinyl radical	1.03	0.89	0.84	0.80	
			one state/irrep		
formaldehyde	0.97	0.87	0.89	0.84	0.72
ethene	1.07	0.95	0.94	0.88	0.83
butadiene	1.12	0.89	0.87	0.81	
acetone	1.00	0.93	0.88	0.76	
vinyl radical	0.87	0.84	0.75	0.78	
	6-31+G ^{*b}		Left Eigenvector		
	6-31++G ^{**}	aug-cc-pVDZ	aug-cc-pVTZ	aug-cc-pVQZ	
			three states/irrep		
formaldehyde	0.97	1.02	1.00	0.99	1.02
ethene	1.02	1.04	1.03	1.03	1.04
butadiene	0.99	1.05	1.01	1.09	
acetone	1.02	1.02	1.05	1.06	
vinyl radical	0.96	1.01	1.00	0.97	
			one state/irrep		
formaldehyde	1.00	0.98	1.02	1.11	0.97
ethene	1.04	1.12	1.09	1.05	1.03
butadiene	0.97	1.05	1.06	0.95	
acetone	0.93	1.03	1.11	1.05	
vinyl radical	1.07	1.00	1.04	1.17	

^aRatio of the number of cycles with the projected amplitude guess with respect to the standard guess. ^bInitial basis set: 6-31G*. Amplitudes are projected from the previous basis in the sequence.

Table 6. Excited State, Second Sequence of Basis Sets^a

	Right Eigenvector			
	three states/irrep		one state/irrep	
	aug-cc-VDZ ^b	d-aug-cc-pVDZ	aug-cc-VDZ ^b	d-aug-cc-pVDZ
formaldehyde	1.03	0.90	0.94	0.85
ethene	0.99	0.93	1.07	0.96
butadiene	0.99	0.81	1.00	0.92
acetone	1.02	0.84	0.92	0.79
vinyl radical	1.02	0.79	0.94	0.67

	Left Eigenvector			
	three states/irrep		one state/irrep	
	aug-cc-VDZ ^b	d-aug-cc-pVDZ	aug-cc-VDZ ^b	d-aug-cc-pVDZ
formaldehyde	0.99	0.93	1.00	0.93
ethene	1.01	0.94	1.05	0.99
butadiene	1.00	0.95	0.95	1.07
acetone	1.00	0.95	1.06	1.05
vinyl radical	0.97	1.07	1.04	0.86

^a Ratio of the number of cycles with the projected amplitude guess with respect to the standard guess. ^b Initial basis set: cc-pVDZ. Amplitudes are projected from the previous basis in the sequence.

in all of the sequences also show that projecting the amplitudes from a basis set without diffuse functions to one with diffuse functions may not be very efficient. On the other hand, a calculation with a basis set with “more” diffuse functions considerably benefits from the projected guess compared to the standard one.

3.3. Excited States. Excited state results are reported in Tables 5–7. Good performance is obtained for the right eigenvectors, although the savings is less than for the ground state *t* amplitudes. The results are better when one state per irrep is considered, while the savings is somewhat less when three states per irrep are computed. This is not surprising, as higher states are more diffuse, and increasing the basis set may lead to larger changes in the wave function. More than for the ground state, the presence of diffuse functions in the basis set used for the projection is important for excited states. Indeed, the CIS eigenvectors solved with the diffuse basis functions are a better guess than projected EOM-CCSD amplitudes without diffuse functions (6-31G* → 6-31+G*, cc-pVDZ → aug-cc-pVDZ, or cc-pVTZ → aug-cc-pVTZ). When a basis is augmented with an additional set of diffuse functions (aug-cc-pVDZ → d-aug-cc-pVDZ or aug-cc-pVTZ → d-aug-cc-pVTZ), the projected guess approach is better than the CIS guess.

For the left eigenvectors, using the converged right eigenvectors (with the same basis set) as the guess is almost consistently better than projecting the left eigenvectors from another basis set, similar to what was found for the ground state *Z* vector.

4. CONCLUSIONS

In this paper, we present a projection technique to use converged CC amplitudes as the initial guess for a new calculation with a different basis set. This technique is based on the corresponding orbitals approach and can be used for ground and excited state calculations. We test a variety of basis sets to define the guess for the next (ground and excited state) calculation on a

Table 7. Excited State, Third Sequence of Basis Sets^a

	Right Eigenvector			
	three states/irrep		one state/irrep	
	aug-cc-VTZ ^b	d-aug-cc-pVTZ	aug-cc-VTZ ^b	d-aug-cc-pVTZ
formaldehyde	0.98	0.84	0.91	0.79
ethene	0.96	0.92	1.03	0.87
vinyl radical	1.01	0.80	0.91	0.58

	Left Eigenvector			
	three states/irrep		one state/irrep	
	aug-cc-VTZ ^b	d-aug-cc-pVTZ	aug-cc-VTZ ^b	d-aug-cc-pVTZ
formaldehyde	0.99	0.93	1.02	0.99
ethene	0.99	0.95	0.98	0.98
vinyl radical	0.94	0.86	1.21	0.87

^a Ratio of the number of cycles with the projected amplitude guess with respect to the standard guess. ^b Initial basis set: cc-pVTZ. Amplitudes are projected from the previous basis in the sequence.

series of small closed and open shell molecules. We also report two examples of calculations with the WIU theory as representative of production level calculations.

The results in this work indicate that the projection strategy is not generally useful to generate an initial guess from another basis set when such result is not already available. In fact, the combined CPU time of a calculation with a smaller basis set and that with a larger basis with the projected guess is always larger than directly performing the calculation with the standard guess. Nevertheless, the projected guess is certainly useful when the calculation with a smaller basis set is already available, which is the case with basis set extrapolation techniques and model chemistry protocols such as W1. In these cases, the projected guess reduces the number of iterations for the solution of the CC equations by 25–30% without adding any significant computational cost. This projection is also very useful in geometry optimizations, especially when approaching the local minimum. For excited states, this methodology is suitable when the interest is in refining the calculation of a particular state by increasing the basis set with more diffuse functions (which are often necessary for accurate results³⁷).

For ground state calculations, the *t* amplitudes benefit from the projected guess, and we find a reduction in the number of cycles from 10% to 50%, while the *Z* amplitudes do not benefit as much. For excited state calculations, the projected guess for the right eigenvectors is considerably better to use than the CIS eigenvectors when additional diffuse functions are included in the basis set. In such a case, the number of iterations for the solution of the EOM-CCSD equations is reduced by 10–20%. On the other hand, for the left eigenvectors, the guess that uses the converged right eigenvectors is very often a better choice than projecting left eigenvectors from a smaller basis set.

We also point out that amplitudes computed from other approximations can be projected as well, for example, projecting CCD or CISD amplitudes as a guess for CCSD, and that looser thresholds for the convergence of the calculation with a smaller basis set can be set if the interest is only in creating the guess. This strategy can even be used in the reverse order, i.e., projecting from a larger to a smaller basis, for example, investigating higher excited states with a less demanding basis set after converging lower ones.

■ ASSOCIATED CONTENT

S Supporting Information. Total number of iterative cycles for the solution of the CCSD and EOM-CCSD equations for all the molecules for each basis set, with the standard and the projected guess. Excitation energies and oscillator strengths for the EOM-CCSD calculations for each basis set. Geometries and energetics for the WIU test molecules. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: marco@gaussian.com.

■ REFERENCES

- (1) Amos, A. T.; Hall, G. G. *Proc. R. Soc. London* **1961**, *263*, 483–493.
- (2) King, H. F.; Stanton, R. E.; Kim, H.; Wyatt, R. E.; Parr, R. G. *J. Chem. Phys.* **1967**, *47*, 1936–1941.
- (3) Bartlett, R. J.; Musial, M. *Rev. Mod. Phys.* **2007**, *79*, 291–352.
- (4) Bartlett, R. J.; Purvis, G. D. *Int. J. Quantum Chem.* **1978**, *14*, 561–581.
- (5) Wiberg, K. B.; de Oliveira, A. E.; Trucks, G. W. *J. Phys. Chem. A* **2002**, *106*, 4192–4199.
- (6) Christiansen, O.; Koch, H.; Jorgensen, P. *Chem. Phys. Lett.* **1995**, *243*, 409–418.
- (7) Koch, H.; Christiansen, O.; Jorgensen, P.; Olsen, J. *Chem. Phys. Lett.* **1995**, *244*, 75–82.
- (8) Schutz, M.; Werner, H. *Chem. Phys. Lett.* **2000**, *318*, 370–378.
- (9) Osted, A.; Kongsted, J.; Mikkelsen, K. V.; Christiansen, O. *J. Phys. Chem. A* **2004**, *108*, 8646–8658.
- (10) Kowalski, K.; Valiev, M. *J. Phys. Chem. A* **2006**, *110*, 13106–13111.
- (11) Caricato, M.; Vreven, T.; Trucks, G. W.; Frisch, M. J.; Wiberg, K. B. *J. Chem. Phys.* **2009**, *131*, 134105.
- (12) Shiozaki, T.; Kamiya, M.; Hirata, S.; Valeev, E. F. *J. Chem. Phys.* **2008**, *129*, 071101.
- (13) Valeev, E. F.; Crawford, T. D. *J. Chem. Phys.* **2008**, *128*, 244113.
- (14) Tew, D. P.; Klopper, W.; Neiss, C.; Haettig, C. *Phys. Chem. Chem. Phys.* **2007**, *9*, 1921–1930.
- (15) Curtiss, L. A.; Raghavachari, K.; Trucks, G. W.; Pople, J. A. *J. Chem. Phys.* **1991**, *94*, 7221–7230.
- (16) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1998**, *109*, 7764–7776.
- (17) Montgomery, J. A., Jr.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A. *J. Chem. Phys.* **1999**, *110*, 2822–2827.
- (18) Martin, J. M. L.; de Oliveira, G. *J. Chem. Phys.* **1999**, *111*, 1843–1856.
- (19) Parthiban, S.; Martin, J. M. L. *J. Chem. Phys.* **2001**, *114*, 6014–6029.
- (20) Tajti, A.; Szalay, P. G.; Császár, A. G.; Kállay, M.; Gauss, J.; Valeev, E. F.; Flowers, B. A.; Vázquez, J.; Stanton, J. F. *J. Chem. Phys.* **2004**, *121*, 11599–11613.
- (21) Pulay, P. *Chem. Phys. Lett.* **1980**, *73*, 393–398.
- (22) Pulay, P. *J. Comput. Chem.* **1982**, *3*, 556–560.
- (23) Scuseria, G. E.; Lee, T. J.; Schaefer, H. F. *Chem. Phys. Lett.* **1986**, *130*, 236–239.
- (24) Purvis, G. D.; Bartlett, R. J. *J. Chem. Phys.* **1981**, *75*, 1284–1292.
- (25) Trucks, G. W.; Noga, J.; Bartlett, R. J. *Chem. Phys. Lett.* **1988**, *145*, 548–554.
- (26) Davidson, E. R. *J. Comput. Phys.* **1975**, *17*, 87–94.
- (27) Hirao, K.; Nakatsuji, H. *J. Comput. Phys.* **1982**, *45*, 246–254.
- (28) Caricato, M.; Trucks, G. W.; Frisch, M. J. *J. Chem. Theory Comput.* **2010**, *6*, 1966–1970.
- (29) Stanton, J. F.; Bartlett, R. J. *J. Chem. Phys.* **1993**, *98*, 7029–7039.
- (30) Handy, N. C.; Pople, J. A.; Head-Gordon, M.; Raghavachari, K.; Trucks, G. W. *Chem. Phys. Lett.* **1989**, *164*, 185–192.
- (31) Kobayashi, R.; Handy, N. C.; Amos, R. D.; Trucks, G. W.; Frisch, M. J.; Pople, J. A. *J. Chem. Phys.* **1991**, *95*, 6723–6733.
- (32) Salter, E. A.; Trucks, G. W.; Bartlett, R. J. *J. Chem. Phys.* **1989**, *90*, 1752–1766.
- (33) Gauss, J.; Stanton, J. F.; Bartlett, R. J. *J. Chem. Phys.* **1991**, *95*, 2623–2638.
- (34) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. In *Numerical Recipes in Fortran*, 2nd ed.; Cambridge University Press: Cambridge, U.K., 1992; pp 51–63.
- (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.; Norm, 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*; Gaussian, Inc.: Wallingford, CT, 2009.
- (36) Barnes, E. C.; Petersson, G. A.; Montgomery, J. A., Jr.; Frisch, M. J.; Martin, J. M. L. *J. Chem. Theory Comput.* **2009**, *5*, 2687–2693.
- (37) Caricato, M.; Trucks, G. W.; Frisch, M. J.; Wiberg, K. B. *J. Chem. Theory Comput.* **2010**, *6*, 370–383.
- (38) Koziol, L.; Levchenko, S. V.; Krylov, A. I. *J. Phys. Chem. A* **2006**, *110*, 2746–2758.
- (39) Barone, V.; Bloino, J.; Biczysko, M. *Phys. Chem. Chem. Phys.* **2010**, *12*, 1092–1101.
- (40) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479–483.
- (41) Chen, W.; Schlegel, H. B. *J. Chem. Phys.* **1994**, *101*, 5957–5968.
- (42) Stanton, J. F. *J. Chem. Phys.* **1994**, *101*, 371–374.

Benchmark Full Configuration Interaction Calculations on the Lowest-Energy 2P and 4P States of the Three-Electron Harmonium Atom

Jerzy Cioslowski* and Eduard Matito*

Institute of Physics, University of Szczecin, Wielkopolska 15, 70-451 Szczecin, Poland

ABSTRACT: Full configuration interaction calculations carried out in conjunction with careful optimization of basis sets and judicious extrapolation schemes for 12 values of the confinement strength ω ranging from 0.1 to 1000.0 provide benchmark energies for the 2P ground state and the 4P first excited state of the three-electron harmonium atom, allowing for numerical verification of the recently obtained second-order energy coefficients and confirming the few available results of Monte Carlo studies. The final energy values, obtained by correcting the extrapolated data for residual errors in the low-order energy coefficients, possess accuracy of ca. 20 μ Hartree for the doublet state and ca. 10 μ Hartree for the quartet one, making them suitable for calibration and testing of approximate electron correlation methods of quantum chemistry. The energy limits for individual angular momenta ranging from 1 to 4 are also available, facilitating comparisons with results of calculations involving finite basis sets. An example of application involving the BLYP and B3LYP functionals is provided.

1. INTRODUCTION

The two-electron harmonium atom, described by the Hamiltonian:

$$\hat{H} = \frac{1}{2} \sum_{i=1}^N (-\nabla_i^2 + \omega^2 r_i^2) + \sum_{i>j=1}^N \frac{1}{r_{ij}} \quad (1)$$

with $N = 2,^{1-3}$ has been repeatedly employed in calibration and benchmarking of approximate electronic structure methods of quantum chemistry, including those based on the density functional theory (DFT).⁴⁻¹² However, although the availability of exact wave functions and energies for certain values of the confinement strength ω greatly facilitates accuracy assessments, the trivial nature of electron correlation in the two-electron species limits its usefulness in such test calculations. In contrast, the three-electron harmonium atom is of potentially greater interest as it allows for infinite tunability of the extent of the correlations between electrons of both same and opposite spins within a single electronic state.

Only very limited data have been accumulated on the three-electron harmonium atom until now. Relatively low-level electronic structure calculations have been carried out, and their results have been compared with the predictions of a pair model valid at the limit of $\omega \rightarrow 0$ ¹³ that has also been investigated from the asymptotic point of view.¹⁴ In addition, a study involving the Hartree–Fock approximation has been published.¹⁵ Thus far, the most accurate energies of several low-energy states of the three-electron harmonium atom have been obtained with Monte Carlo calculations for three values of ω , namely 1/100, 1/2, and 10.¹⁶

In a recent pilot study, the ground-state energies of the two-electron harmonium atom have been computed with a few μ Hartree accuracy for 20 values of ω ranging from 0.03 to 1000.¹⁷ The full configuration interaction (FCI) approach has been used in conjunction with even-tempered Gaussian basis sets

and judicious extrapolations to both the individual angular momentum and the complete basis set (CBS) limits. As basis sets with only few angular momenta are employed in actual calculations, the availability of both these limits is of crucial importance to testing of approximate approaches of the electronic structure theory.

Encouraged by the performance of this computational scheme, we have recently performed analogous benchmark calculations for the two lowest-energy states of the three-electron harmonium atom. Results of these calculations are presented in this paper.

2. DETAILS OF CALCULATIONS

The calculations described here have been carried out for the 2P and 4P lowest-energy states of the three-electron harmonium atom that arise, respectively, from the $s\bar{s}p_z$ and $sp_x p_y$ configurations of the weak-correlation limit. The FCI energies have been computed for 12 values of ω , namely 0.1, 0.15, 0.2, 0.3, 0.4, 0.5, 1.0, 2.0, 5.0, 10.0, 100.0, and 1000.0, with a modified program of Knowles and Handy.¹⁸ The respective one- and two-electron integrals, generated with the Gaussian09 suite of programs,¹⁹ have involved uncontracted basis sets that, for each value of the angular momentum between 0 and L ($1 \leq L \leq 4$), comprise equal numbers N ($4 \leq N \leq 8$) of spherical Gaussian primitives with exponents $\zeta_{L,N}^k(\omega)$ even-tempered²⁰ according to the formula:

$$\zeta_{L,N}^k(\omega) = \frac{\omega}{2} \alpha_{L,N}(\omega) [\beta_{L,N}(\omega)]^{k-1}, \quad 1 \leq k \leq N \quad (2)$$

While depending on N , the maximum angular momentum L and the electronic state in question, the optimized parameters

Received: November 11, 2010

Published: March 09, 2011

Table 1. Optimized Basis Set Parameters $\alpha_{L,N}(\omega)$ and $\beta_{L,N}(\omega)$ [eq 2] and Corresponding FCI energies $E_{L,N}(\omega)$ of the ^2P State of the Three-Electron Harmonium Atom with $\omega = 1/2$

N	$\alpha_{1,N}(1/2)$	$\beta_{1,N}(1/2)$	$E_{1,N}(1/2)$	$\alpha_{2,N}(1/2)$	$\beta_{2,N}(1/2)$	$E_{2,N}(1/2)$	$\alpha_{3,N}(1/2)$	$\beta_{3,N}(1/2)$	$E_{3,N}(1/2)$
4	0.916687	1.546208	4.0436825	0.918520	1.559219	4.0169503	0.916222	1.588256	4.0143476
5	0.923655	1.502709	4.0436488	0.925382	1.512923	4.0169019	0.923432	1.536571	4.0142866
6	0.929189	1.466172	4.0436378	0.930788	1.475097	4.0168859	0.929089	1.494644	4.0142661
7	0.933634	1.435840	4.0436336	0.935130	1.443885	4.0168797	0.933653	1.460256	4.0142580
8	0.937386	1.410345	4.0436318	0.938764	1.417514	4.0168770	0.937502	1.431465	4.0142544
∞			4.0436301			4.0168744			4.0142510

$\alpha_{L,N}(\omega)$ and $\beta_{L,N}(\omega)$ that minimize the FCI energies $E_{L,N}(\omega)$ have been kept equal for individual Gaussian primitives irrespective of their angular momenta. The quantities $-\ln[-N\omega^{1/2} \ln \alpha_{L,N}(\omega)]$ and $N \ln \beta_{L,N}(\omega)$ have been found accurately approximated by quadratic polynomials in $\omega^{-1/2}$, which aids in the choice of the initial values for $\alpha_{L,N}(\omega)$ and $\beta_{L,N}(\omega)$ and speeds up the energy minimizations.

For $1 \leq L \leq 3$, the previously employed extrapolations to the $N \rightarrow \infty$ limits $E_L(\omega)$, based upon fitting of the parameters of the approximate equation:¹⁷

$$E_{L,N}(\omega) = E_L(\omega) + A_{1,L}(\omega)e^{-\lambda_{1,L}(\omega)N} + A_{2,L}(\omega)e^{-\lambda_{2,L}(\omega)N} \quad (3)$$

have been used. However, for $L = 4$, an alternative extrapolation scheme:

$$E_4(\omega) = E_{4,5}(\omega) + \frac{E_{4,5}(\omega) - E_{4,4}(\omega)}{E_{3,5}(\omega) - E_{3,4}(\omega)} [E_3(\omega) - E_{3,5}(\omega)] \quad (4)$$

which does not require the values of $E_{4,6}(\omega)$, $E_{4,7}(\omega)$, and $E_{4,8}(\omega)$, has been found equally accurate. Consequently, it has been adopted in the actual calculations, allowing for significant reduction of computational effort. For $1 \leq L \leq 3$, the extrapolated energies are lower than their counterparts computed with 8 primitives by at most 9.2 and 0.2 $\mu\text{Hartree}$ (for the ^2P and ^4P states, respectively). For $L = 4$, the extrapolation (4) results in the lowerings of 111.0 and 5.1 $\mu\text{Hartree}$ with respect to the $E_{4,5}(\omega)$ energies.

The estimates $E_L(\omega)$ (that comprise the energy contributions of partial waves with angular momenta up to L) have been extrapolated to the respective CBS limits $E(\omega)$ by fitting the values of $E_L(\omega)$ for $L = 2, 3$, and 4 to the expression:^{21–23}

$$E_L(\omega) = E(\omega) + \frac{B(\omega)}{[L + C(\omega)]^\lambda} \quad (5)$$

where the exponent λ depends on the state in question ($\lambda = 3$ for ^2P and $\lambda = 5$ for ^4P). Extrapolations based upon the values of $E_L(\omega)$ for $L = 1, 2$, and 3 have turned out to be of significantly inferior accuracy and thus have not been used.

For each of the two states under study, the energies $E_L(\omega)$ computed for 12 values of ω have been fitted to the truncated power series:

$$E_L(\omega) = \sum_{k=0}^6 E_L^{(k)} \omega^{(2-k)/2} \quad (6)$$

pertinent to the large ω asymptotics²⁴ with the energy coefficients $E_L^{(0)}$ and $E_L^{(1)}$ kept fixed at their exact L independent values (see below). The same fitting has been carried out for the

estimated CBS limits $E(\omega)$, producing the coefficients $E^{(k)}$ ($2 \leq k \leq 6$). Inspection of the resulting values of $E^{(2)}$ has revealed small but significant deviations from their recently available²⁴ exact counterparts. Consequently, the final energies $E(\omega)$ have been recomputed from the series (6) with corrected coefficients $E^{(k)}$ obtained as follows: For $0 \leq k \leq 2$, the exact values have been used, whereas for $3 \leq k \leq 6$, the values extrapolated from the coefficients $E_L^{(k)}$ ($2 \leq L \leq 4$) and $E^{(k)}$ have been employed, the pertinent extrapolations assuming the deviations of the higher-order coefficients from their exact values being given by quadratic polynomials of the respective deviations of the computed second-order coefficients. For the ^2P state, this final correction lowers the energies by as much as 66 $\mu\text{Hartree}$ (e.g., by 55 $\mu\text{Hartree}$ for $\omega = 10$ and 22 $\mu\text{Hartree}$ for $\omega = 1/2$). The corrections are much smaller for the ^4P state, amounting to less than 5 $\mu\text{Hartree}$ (e.g., 3.3 $\mu\text{Hartree}$ for $\omega = 10$ and 4.6 $\mu\text{Hartree}$ for $\omega = 1/2$).

3. RESULTS AND DISCUSSION

The ^2P doublet ground state arises from the $s\bar{s}p_z$ configuration that is dominant at the small-correlation (large ω) limit with $E^{(0)} = 11/2$, $E^{(1)} = (5/2)(2/\pi)^{1/2}$, and the exact second-order energy coefficient that reads:²⁴

$$E^{(2)} = \frac{49}{9} + \frac{1}{6\pi} [-88 + 2\sqrt{3} - 173 \ln 2 + 98 \ln(1 + \sqrt{3})] \approx -0.176654 \quad (7)$$

The optimized basis-set parameters $\alpha_{L,N}(\omega)$ and $\beta_{L,N}(\omega)$ [eq 2] and the corresponding FCI energies $E_{L,N}(\omega)$ computed for $\omega = 1/2$ that are compiled in Table 1 exhibit smooth convergence to the $N \rightarrow \infty$ limits. Although this convergence is observed for all the aforementioned values of the confinement strength, linear dependencies among the basis functions limit the present calculations to $\omega \geq 0.1$ (compared with $\omega \geq 0.03$ in the case of the two-electron harmonium atom).¹⁷

The energy limits $E_L(\omega)$ for individual angular momenta are listed in Table 2 together with the extrapolated and corrected CBS limits $E(\omega)$. For the two values of ω , namely 10 and 1/2, for which the results of Monte Carlo calculations are available,¹⁶ the corrected energies are found to be in excellent agreement with the previously published data. The computed energy coefficients (Table 3) are of potential use as benchmarks in checking and debugging of future analytical work. The differences between the extrapolated and the corrected values of these coefficients are quite significant, amounting to 68 $\mu\text{Hartree}$ for $E^{(2)}$ and 43 $\mu\text{Hartree}$ for $E^{(3)}$.

The interpolation of $E(\omega)$ between the large ω limit (eq 6) and the small ω limit of

$$E(\omega) = \tilde{E}^{(0)} \omega^{2/3} + \tilde{E}^{(1)} \omega \quad (8)$$

Table 2. Partial Wave and CBS Limits of the ^2P State Energies of the Three-Electron Harmonium Atom

ω	$E_1(\omega)$	$E_2(\omega)$	$E_3(\omega)$	$E_4(\omega)$	$E(\omega)^a$	$E(\omega)^b$
1000.	5562.944818	5562.910896	5562.905536	5562.903916	5562.902484	5562.902417
100.	569.814500	569.780963	569.775796	569.774252	569.772901	569.772838
10.	61.177964	61.145631	61.141032	61.139708	61.138588 ^c	61.138533 ^d
5.	31.832079	31.800462	31.796175	31.794970	31.793973	31.793923
2.	13.695898	13.665666	13.661932	13.660935	13.660149	13.660107
1.	7.373142	7.344405	7.341195	7.340394	7.339798	7.339766
0.5	4.043630	4.016874	4.014251	4.013669	4.013274 ^c	4.013253 ^d
0.4	3.346560	3.320557	3.318126	3.317615	3.317284	3.317266
0.3	2.630164	2.605217	2.603028	2.602608	2.602352	2.602340
0.2	1.883885	1.860586	1.858712	1.858413	1.858251	1.858246
0.15	1.492734	1.470709	1.469030	1.468807	1.468698	1.468699
0.1	1.081236	1.061133	1.059671	1.059538	1.059485	1.059493

^a From eq 5 with $\lambda = 3$. ^b After the final correction (see the text for explanation). ^c Compare with the energies of 61.138525, 61.138549, and 61.139485 obtained from the Monte Carlo calculations. ^d Compare with the energies of 4.013240, 4.013224, and 4.013511 obtained from the Monte Carlo calculations.¹⁶

Table 3. Partial Wave and CBS Limits of the Energy Coefficients $E^{(k)}$ [eq 6] for the ^2P State of the Three-Electron Harmonium Atom

	$L = 1^a$	$L = 2^a$	$L = 3^a$	$L = 4^a$	extrapolated ^b	corrected ^c
$E^{(2)}$	-1.33904×10^{-1}	-1.68007×10^{-1}	-1.73460×10^{-1}	-1.75115×10^{-1}	-1.76586×10^{-1}	-1.76654×10^{-1}
$E^{(3)}$	1.29505×10^{-2}	1.86390×10^{-2}	2.15854×10^{-2}	2.27364×10^{-2}	2.39767×10^{-2}	2.40199×10^{-2}
$E^{(4)}$	-5.79129×10^{-4}	-8.26331×10^{-4}	-1.62200×10^{-3}	-1.96851×10^{-3}	-2.40436×10^{-3}	-2.41347×10^{-3}
$E^{(5)}$	-4.11410×10^{-5}	-1.32704×10^{-4}	-3.75221×10^{-5}	1.69114×10^{-5}	9.36887×10^{-5}	9.49301×10^{-5}
$E^{(6)}$	4.95919×10^{-6}	1.87513×10^{-5}	1.49576×10^{-5}	1.12262×10^{-5}	5.48519×10^{-6}	5.39176×10^{-6}

^a Obtained by fitting of the extrapolated energies $E_L(\omega)$ to eq 6. ^b Obtained by fitting of the extrapolated energies $E(\omega)$ to eq 6. ^c After the final correction (see the text for explanation). For $E^{(2)}$, the exact value (eq 7) is listed.

Table 4. Coefficients of the Padé Approximant (eq 9) for the Energies of the ^2P State of the Three-Electron Harmonium Atom

k	a_k	b_k	c_k	d_k
1	3.82211×10^{-4}	2.66583×10^{-4}	-1.03821	3.72579×10^{-1}
2	-8.22620×10^{-3}	1.17287×10^{-2}	-6.62884×10^{-1}	2.72429×10^{-1}
3	-2.65714×10^{-2}	1.90218×10^{-2}	-3.31207×10^{-1}	2.31675×10^{-1}
4	7.26721×10^{-3}	-7.89154×10^{-4}	2.20583×10^{-1}	1.44545×10^{-1}
5	-1.05700×10^{-2}	-3.50039×10^{-3}	6.26591×10^{-1}	1.62447×10^{-1}
6	3.77181×10^{-2}	1.82252×10^{-2}	9.57521×10^{-1}	2.34292×10^{-1}

where $\tilde{E}^{(0)} = (1/2)3^{5/3}$ and $\tilde{E}^{(1)} = (1/2)(3 + 3^{1/2} + 6^{1/2})^{14}$ can be readily accomplished with the help of the Padé approximants using $\omega^{1/6}$ as their arguments.²⁴ In particular, requesting the expression in which $\omega^{2/3}$ multiplies the [14/12] approximant to conform to both the limits (6) and (8) (and noting that the coefficients that multiply the $\omega^{-13/6}$ and $\omega^{-7/3}$ terms in the large ω power series vanish together with the small ω ones multiplying the $\omega^{7/6}$ and $\omega^{3/2}$ terms) uniquely determines all but one of the pertinent 27 coefficients. In turn, minimization of the maximum error with respect to the remaining coefficient yields the approximate expression:

$$E(\omega) = E^{(0)}\omega + E^{(1)}\omega^{1/2} + E^{(2)} + \sum_{k=1}^6 \frac{a_k \omega^{1/6} + b_k}{\omega^{1/3} + c_k \omega^{1/6} + d_k} \quad (9)$$

(see Table 4 for the values of a_k , b_k , c_k , and d_k) that reproduces the corrected energies $E(\omega)$ within 5.6 $\mu\text{Hartree}$. Interestingly, eq 9

affords the estimate $E(1/100) = 0.181677$ that compares quite well with the Monte Carlo result of $E(1/100) = 0.181936$.¹⁶ In contrast, the power series (6) produces the poor estimate of $E(1/100) = 0.225517$.

The data computed for the ^4P quartet lowest-energy excited state that arises from the sp_xp_y configuration are presented in Tables 5–8. In this case, $E^{(0)} = 13/2$, $E^{(1)} = 2(2/\pi)^{1/2}$, and

$$E^{(2)} = \frac{23}{9} + \frac{8}{3\pi} [-4 + \sqrt{3} - 7 \ln 2 + 4 \ln(1 + \sqrt{3})] \approx -0.0756103 \quad (10)$$

whereas the coefficients $E^{(0)}$ and $E^{(1)}$ are the same as those pertaining to the ^2P state.^{14,24} The reduced electron–electron repulsion, reflected in the smaller absolute values of $E^{(1)}$ and $E^{(2)}$, results in higher convergence rates of the FCI energies. This enhanced convergence is apparent in both the dependences of

Table 5. Optimized Basis Set Parameters $\alpha_{L,N}(\omega)$ and $\beta_{L,N}(\omega)$ [eq 2] and Corresponding FCI Energies $E_{L,N}(\omega)$ of the 4P State of the Three-Electron Harmonium Atom with $\omega = 1/2$

N	$\alpha_{1,N}(1/2)$	$\beta_{1,N}(1/2)$	$E_{1,N}(1/2)$	$\alpha_{2,N}(1/2)$	$\beta_{2,N}(1/2)$	$E_{2,N}(1/2)$	$\alpha_{3,N}(1/2)$	$\beta_{3,N}(1/2)$	$E_{3,N}(1/2)$
4	0.942768	1.434508	4.3259606	0.943263	1.442992	4.3125639	0.941614	1.465746	4.3109809
5	0.947999	1.399152	4.3259583	0.948426	1.404443	4.3125593	0.947005	1.423446	4.3109748
6	0.951595	1.373620	4.3259577	0.952196	1.376518	4.3125582	0.950936	1.392373	4.3109732
7	0.954431	1.353081	4.3259576	0.955084	1.354419	4.3125579	0.953982	1.367908	4.3109728
8	0.956760	1.335578	4.3259575	0.957437	1.336069	4.3125578	0.956436	1.348491	4.3109726
∞			4.3259575			4.3125577			4.3109725

Table 6. Partial Wave and CBS Limits of the 4P State Energies of the Three-Electron Harmonium Atom

ω	$E_1(\omega)$	$E_2(\omega)$	$E_3(\omega)$	$E_4(\omega)$	$E(\omega)^a$	$E(\omega)^b$
1000.	6550.405429	6550.389639	6550.387757	6550.387399	6550.387238	6550.387236
100.	665.900725	665.885057	665.883202	665.882854	665.882700	665.882697
10.	69.990094	69.974811	69.973031	69.972712	69.972576 ^c	69.972573 ^c
5.	36.012550	35.997498	35.995757	35.995455	35.995329	35.995325
2.	15.201977	15.187380	15.185701	15.185429	15.185322	15.185318
1.	8.041988	8.027896	8.026268	8.026027	8.025938	8.025933
0.5	4.325958	4.312558	4.310973	4.310770	4.310703 ^d	4.310698 ^d
0.4	3.557361	3.544232	3.542657	3.542469	3.542409	3.542404
0.3	2.772905	2.760165	2.758597	2.758428	2.758378	2.758374
0.2	1.963740	1.951627	1.950059	1.949919	1.949882	1.949879
0.15	1.544118	1.532508	1.530930	1.530812	1.530784	1.530782
0.1	1.107280	1.096459	1.094854	1.094768	1.094750	1.094751

^a From eq 5 with $\lambda = 5$. ^b After the final correction (see the text for explanation). ^c Compare with the energies of 69.972571, 69.972571, and 69.972624 obtained from the Monte Carlo calculations. ^d Compare with the energies of 4.310690, 4.310690, and 4.310712 obtained from the Monte Carlo calculations.¹⁶

Table 7. Partial Wave and CBS Limits of the Energy Coefficients $E^{(k)}$ [eq 6] for the 4P State Energies of the Three-Electron Harmonium Atom

	$L = 1^a$	$L = 2^a$	$L = 3^a$	$L = 4^a$	extrapolated ^b	corrected ^c
$E^{(2)}$	-5.73403×10^{-2}	-7.31870×10^{-2}	-7.50824×10^{-2}	-7.54447×10^{-2}	-7.56081×10^{-2}	-7.56103×10^{-2}
$E^{(3)}$	3.76541×10^{-3}	5.55339×10^{-3}	5.97127×10^{-3}	6.11889×10^{-3}	6.21146×10^{-3}	6.20753×10^{-3}
$E^{(4)}$	-2.12366×10^{-4}	-2.25364×10^{-4}	-4.04234×10^{-4}	-4.35596×10^{-4}	-4.55307×10^{-4}	-4.53956×10^{-4}
$E^{(5)}$	6.32038×10^{-6}	-1.66290×10^{-5}	1.32753×10^{-5}	1.89166×10^{-5}	2.05880×10^{-5}	2.08485×10^{-5}
$E^{(6)}$	1.02964×10^{-7}	2.37962×10^{-6}	4.94165×10^{-7}	-6.06493×10^{-8}	-8.38971×10^{-8}	-1.52385×10^{-7}

^a Obtained by fitting of the extrapolated energies $E_L(\omega)$ to eq 6. ^b Obtained by fitting of the extrapolated energies $E(\omega)$ to eq 6. ^c After the final correction (see the text for explanation). For $E^{(2)}$, the exact value (eq 10) is used.

Table 8. Coefficients of the Padé Approximant (eq 9) for the Energies of the 4P State of the Three-Electron Harmonium Atom

k	a_k	b_k	c_k	d_k
1	1.43587×10^{-2}	-1.61112×10^{-3}	2.56382×10^{-1}	-4.13474×10^{-2}
2	-5.96215×10^{-5}	8.18610×10^{-4}	-7.25430×10^{-1}	2.42079×10^{-1}
3	-1.17585×10^{-2}	8.50060×10^{-3}	-3.78158×10^{-1}	1.77911×10^{-1}
4	-8.28147×10^{-6}	2.65891×10^{-5}	-2.32442×10^{-1}	4.67866×10^{-1}
5	1.25587×10^{-3}	-1.05411×10^{-4}	1.41377×10^{-1}	5.89617×10^{-2}
6	-3.78817×10^{-3}	-2.04825×10^{-3}	7.31118×10^{-1}	1.60196×10^{-1}

$E_{L,N}(\omega)$ on N (Table 5) and $E_L(\omega)$ on L (Table 6). Consequently, the agreement among the extrapolated and corrected energies and those published previously¹⁶ is even closer than in the case of the 2P state. The convergence of the energy coefficients to their CBS limits is also noticeably faster, the differences between the extrapolated and corrected values of

$E^{(2)}$ and $E^{(3)}$ amounting to 2.2 and 3.9 μ Hartree, respectively (Table 7).

The quality of the Padé approximation for the energies of the 4P state is mixed. On one hand, the approximant with the coefficients listed in Table 8 reproduces these energies within 0.7 μ Hartree and yields the estimate of $E(1/100) = 0.182844$ that

compares well with the Monte Carlo result of $E(1/100) = 0.182973$ [although the improvement over the power-series result of $E(1/100) = 0.184971$ is not as dramatic as for the analogous ^2P energy]. On the other hand, the approximant possesses a pole at $\omega \approx 2 \times 10^{-6}$ that precludes smooth interpolation between the small- and large-correlation limits.

4. EXAMPLE OF APPLICATION

When employed in conjunction with the Kohn–Sham formalism, the BLYP^{25,26} and B3LYP²⁷ functionals allow for inclusion of electron correlation effects within the framework of one-electron theory, yielding reasonably accurate energies at relatively low computational cost. For this reason, these functionals remain one of the most popular choices among quantum chemists carrying out calculations on medium- and large-size systems.²⁸

It is instructive to compare the performance of the BLYP and B3LYP functionals for the two lowest-energy electronic states of the three-electron harmonium atom. The accuracy of the approximate DFT energies is readily assessed by plotting $\Delta E(\omega) = E(\omega) - E^{(0)} - \omega - E^{(1)}\omega^{1/2}$ (the energy contributions beyond the first order) vs $\omega^{-1/2}$. Inspection of Figure 1, in which the relevant HF energy contributions are also displayed, reveals relatively minor variations of the exact and the HF values of $\Delta E(\omega)$ with ω , their difference (which amounts to the correlation energy) depending weakly on ω . In contrast, the BLYP and B3LYP functionals fail spectacularly at the weak-correlation limit, the computed values of $\Delta E(\omega)$ exhibiting divergencies as ω tends to infinity. These divergencies are caused by the inaccurate exchange components of the approximate total energies.

For the ^2P doublet ground state of the three-electron harmonium atom, the performance of the B3LYP functional becomes quite satisfactory for $\omega \leq 2$ and does not worsen even at $\omega = 0.1$, which corresponds to the weakest confinement for which the present data are available. In contrast, the BLYP functional accounts for only a fraction of the electron correlation energy for smaller values of ω while sharing the large ω singular behavior with its B3LYP counterpart. The B3LYP energies of the ^4P quartet state are not as accurate as those of the doublet state but still follow the exact ones quite closely. Interestingly, the BLYP energies appear to converge to their HF rather than exact counterparts as ω decreases.

Overall, as expected, this test clearly favors the B3LYP functional over its BLYP congener. It also uncovers the need for improvement in the handling of the exchange part in these functionals.

It should be emphasized that the above picture of (dis)agreements among approximate energies computed at different levels of theory does not persist for smaller values of ω (and specially at the $\omega \rightarrow 0$ limit). Upon weakening of the confinement, localization of electrons (analogous to the Wigner crystallization of the homogeneous electron gas) occurs,^{4,29} bringing about breakdown of the predominantly single-determinantal nature of the electronic wave function that manifests itself in the asymptotic vanishing of all natural orbital occupancies.³⁰

5. CONCLUSIONS

Full configuration interaction calculations carried out in conjunction with careful optimization of basis sets and judicious extrapolation schemes provide benchmark energies for the ^2P ground state and the ^4P first excited state of the three-electron harmonium atom, allowing for numerical verification of the

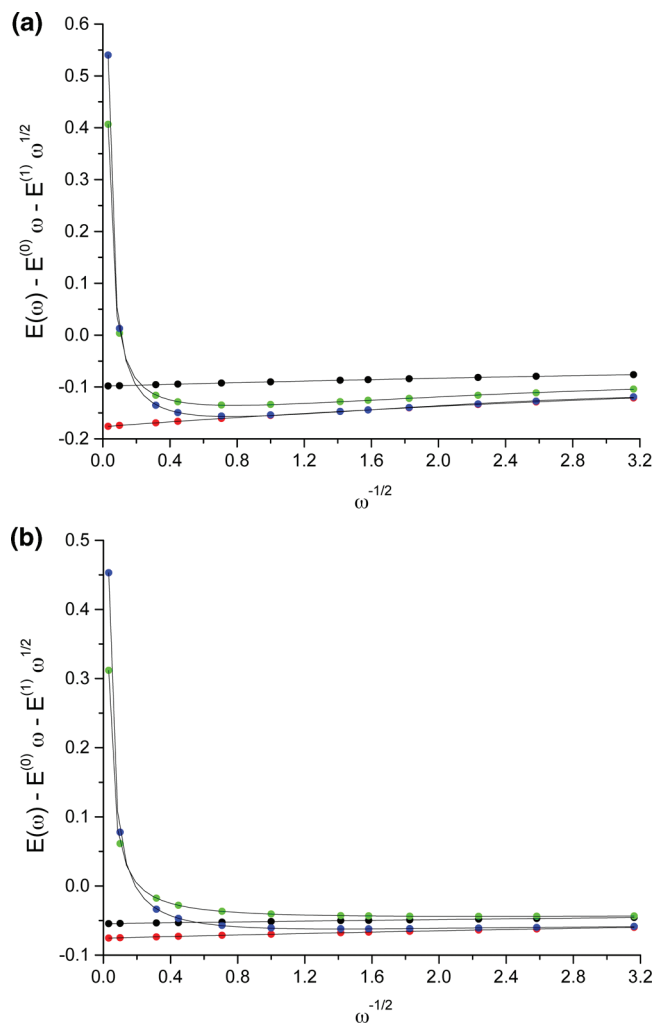


Figure 1. The energy contributions beyond the first order vs $\omega^{-1/2}$ for: (a) the ^2P state and (b) the ^4P state of the three-electron harmonium atom (red: exact; black: HF; green: BLYP; and blue: B3LYP).

recently obtained second-order energy coefficients²⁴ and confirming the few available results of Monte Carlo studies.¹⁶ The final energy values, obtained by correcting the extrapolated data for residual errors in the low-order energy coefficients, appear to possess accuracy of ca. $20 \mu\text{Hartree}$ for the doublet state and ca. $10 \mu\text{Hartree}$ for the quartet one, making them suitable for calibration and testing of approximate electron correlation methods of quantum chemistry. The energy limits for individual angular momenta ranging from 1 to 4 are also available, facilitating comparisons with results of calculations involving finite basis sets.

For confinement strengths other than those studied in this paper, the energies can be readily computed with sufficient accuracy from the respective Padé approximants. If needed, the corresponding wave functions can also be obtained by following the present computational procedure while taking advantage of the aforescribed dependence of the parameters $\alpha_{L,N}(\omega)$ and $\beta_{L,N}(\omega)$ on ω that provides a prescription for rapid construction of the optimal basis sets.

Extension of the present work to higher excited states of the three-electron harmonium atoms and to species with four and five electrons is conceptually straightforward and will be carried out in the near future.

AUTHOR INFORMATION

Corresponding Authors

*E-mail: jerzy@wmf.univ.szczecin.pl; ematito@gmail.com.

ACKNOWLEDGMENT

This work has been supported by Marie Curie IntraEuropean Fellowship, Seventh Framework Programme (FP7/2007-2013), under grant agreement PIEF-GA-2008-221734 (E.M.) and the Polish Ministry of Science and Higher Education (project no. N N204 215634).

REFERENCES

- (1) Taut, M. *Phys. Rev. A: At., Mol., Opt. Phys.* **1993**, *48*, 3561.
- (2) Taut, M. *J. Phys. A: Math. Gen.* **1994**, *27*, 1045.
- (3) Cioslowski, J.; Pernal, K. *J. Chem. Phys.* **2000**, *113*, 8434 and the references cited therein.
- (4) Laufer, P. M.; Krieger, J. B. *Phys. Rev. A: At., Mol., Opt. Phys.* **1986**, *33*, 1480–1491.
- (5) Kais, S.; Hersbach, D. R.; Handy, N. C.; Murray, C. W.; Laming, G. J. *J. Chem. Phys.* **1993**, *99*, 417.
- (6) Qian, Z.; Sahni, V. *Phys. Rev. A: At., Mol., Opt. Phys.* **1998**, *57*, 2527.
- (7) Filippi, C.; Umrigar, C. J.; Taut, M. *J. Chem. Phys.* **1994**, *100*, 1290.
- (8) Taut, M.; Ernst, A.; Eschrig, H. *J. Phys. B: At. Mol. Phys.* **1998**, *31*, 2689.
- (9) Huang, C.-J.; Umrigar, C. J. *Phys. Rev. A: At., Mol., Opt. Phys.* **1997**, *56*, 290.
- (10) Hessler, P.; Park, J.; Burke, K. *Phys. Rev. Lett.* **1999**, *82*, 378.
- (11) Ivanov, S.; Burke, K.; Levy, M. *J. Chem. Phys.* **1999**, *110*, 10262.
- (12) Zhu, W. M.; Trickey, S. B. *J. Chem. Phys.* **2006**, *125*, 094317.
- (13) Taut, M.; Pernal, K.; Cioslowski, J.; Staemmler, V. *J. Chem. Phys.* **2003**, *118*, 4818.
- (14) Cioslowski, J.; Pernal, K. *J. Chem. Phys.* **2006**, *125*, 064106.
- (15) Sundqvist, P. A.; Volkov, S. Y.; Lozovik, Y. E.; Willander, M. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2002**, *66*, 075335.
- (16) Varga, K.; Navratil, P.; Usukura, J.; Suzuki, Y. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2001**, *63*, 205308.
- (17) Matito, E.; Cioslowski, J.; Vyboishchikov, S. F. *Phys. Chem. Chem. Phys.* **2010**, *12*, 6712.
- (18) Knowles, P.; Handy, N. *Comput. Phys. Commun.* **1986**, *54*, 75.
- (19) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; R. L. Martin, K. Morokuma, Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Ö. Farkas, Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, revision A.02, Gaussian, Inc.: Wallingford, CT, 2009.
- (20) Ruedenberg, K.; Raffanetti, R. C.; Bardo, R. D. Energy, Structure, and Reactivity. *Proceedings of the 1972 Boulder Seminar Research Conference on Theoretical Chemistry*; Smith, D. W., Ed.; Wiley: New York, 1973; p 164.
- (21) Hill, R. N. *J. Chem. Phys.* **1985**, *83*, 1173.
- (22) Kutzelnigg, W.; Morgan, J. D. *J. Chem. Phys.* **1992**, *96*, 4484.
- (23) Kutzelnigg, W.; Morgan, J. D. *J. Chem. Phys.* **1992**, *97*, 8821 (Erratum).
- (24) Cioslowski, J.; Matito, E. *J. Chem. Phys.*; in press, DOI:10.1063/1.3553558.
- (25) Becke, A. *Phys. Rev. A: At., Mol., Opt. Phys.* **1988**, *38*, 3098.
- (26) Lee, C.; Wang, Y.; R. P. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785.
- (27) Becke, A. *J. Chem. Phys.* **1993**, *98*, 5648.
- (28) Swart, M. personal communication.
- (29) Cioslowski, J.; Grzebielucha, E. *Phys. Rev. A: At., Mol., Opt. Phys.* **2008**, *77*, 032508.
- (30) Cioslowski, J.; Buchowiecki, M. *J. Chem. Phys.* **2006**, *125*, 064105.

Distributed Multipoles and Energies of Flexible Molecules

Hai-Anh Le and Ryan P. A. Bettens*

Department of Chemistry, National University of Singapore, 3 Science Drive 3, Singapore 117543

S Supporting Information

ABSTRACT: In this work we show that energies and distributed multipoles, up to and including rank two, can be accurately determined via a modified Shepard interpolation of ab initio data for small molecules. The molecules considered here are the amino aldehydes, Gly and Ala, which may be typical smaller fragment molecules in certain molecular energy-based fragmentation schemes. The method is general and should be suitable for applications also involving crystal structure prediction, modeling molecular clusters, and Monte Carlo or molecular/reaction dynamics simulations. The configuration space covered by the interpolation includes that sampled by the Gly and Ala peptides in protein crystal structures, i.e., 12 dimensions for Gly: 3 torsion angles (φ , ψ , ω), 5 bond lengths, and 4 bond angles and 15 dimensions for Ala: 4 torsion angles, 6 bond lengths, and 5 bond angles. In this work we also describe a new method of importance sampling the relevant configuration spaces, and show that it is possible to interpolate “axis free” multipoles.

1. INTRODUCTION

It is now well established that an inexpensive but accurate approach to determine the electrostatic interaction energy between molecules, or the electrostatic potential about them, is by distributing multipoles at various sites within molecules. Different methods to determine the multipoles are available and include “distributed multipole analysis” (DMA),^{1,2} “atoms in molecules” (AIM)³ and “transferable atom equivalent” (TAE),⁴ and “cumulative atomic multipole moments” (Camm).^{5,6}

All of these approaches utilize electronic structure calculations to obtain the distributed multipoles. Necessarily a specific molecular structure must be adopted. The molecular structure is precisely defined in terms of the atomic configuration that can be represented by an appropriate set of internal coordinates. Thus for a specific electronic structure calculation the distributed multipoles are only exact for the specific point in molecular configuration space where the calculation was performed. This is all that is required for applications involving rigid molecules. However, for systems involving flexible molecules different parts of configuration space are visited, so the distributed multipoles that were determined at a single point in this space are no longer exact. One possible option is to assume that the electron density remains unchanged in different molecular conformations. Unfortunately such an assumption is not sufficient for applications involving the modeling of molecular clusters⁷ or crystal structure prediction.⁸ Aside from performing a new electronic structure calculation at the new molecular configuration, which at the very least is incredibly expensive or at worst impossible depending upon the application, various attempts have been made at predicting the distributed multipoles.

One such attempt includes substantially reducing the dimensionality of configuration space to focus only upon a few degrees of freedom, namely selected torsions, then to either parametrically fit the multipoles to ad hoc functions in the reduced space^{7,9} or by interpolation using a grid of precomputed configurations in this reduced space.^{10,11} Both of these types of approaches, while

very satisfactory for their desired application, are only practical provided the dimensionality of the space remains small. For substantially higher dimensionality neither of these approaches can be adopted. Alternatively, the change in the distributed multipoles with conformation has been dealt with using “intramolecular polarization”^{12,13} and then utilized in molecular dynamics.¹⁴

Here we focus our attention on DMA and present a method that enables accurate predictions of both distributed multipoles and intramolecular electronic energies for generally flexible molecules, i.e., molecules in any arbitrary configuration. Our method utilizes the modified Shepard interpolation.^{15–18} It should be noted that integral to this method is the importance sampling of the relevant configuration space to the application of interest, which results in a significant improvement in computational efficiency. Such a method can readily find application in crystal structure prediction, modeling molecular clusters, and Monte Carlo or molecular/reaction dynamics simulations or in an ab initio molecular database.¹⁹ However, depending on the specific application, the resulting sampled surfaces may not be sufficiently accurate if they are directly transferred to a different application. In the latter case the method can still be applied, but the surface should be importance sampled again on the relevant regions of configuration space. The application of most interest to us here is molecular energy-based fragmentation. This particular application is an approximately linear scaling technique that requires the energies, and other properties like the electrostatic potential, of relatively small molecules in a range of configurations. A number of groups have developed various molecular energy-based fragmentation methods by fragmenting a larger molecule, e.g., a protein, into many smaller complete molecules then linearly combining their electronic energies to approximate the total energy, or some other property, of the target molecule.

Received: November 25, 2010

Published: March 18, 2011

The first type of molecular energy-based fragmentation was attempted by Gadre's group using their molecular tailoring method.^{20–24} This method was originally suggested to calculate one-electron properties, such as the electrostatic potential. More recently Zhang and Zhang developed a quite different fragmentation algorithm²⁵ (molecular fractionation with conjugated caps) and have applied (and extended) their approach successfully to several systems.^{26–39} This method was originally designed to accurately compute interaction energies between two molecular systems. Molecular total energies were first attempted using molecular energy-based fragmentation by Li's group^{40–46} (from which we have borrowed the term molecular energy-based fragmentation) and the Collins group^{47–50} and later by ourselves.^{51–54} It should be noted here that other types of linear scaling methods exist, e.g., the density matrix divide and conquer approach of Yang and Lee⁵⁵ or Kitaura's fragment molecular orbital approach (ref 56 and recently ref 57 and references therein), which has been implemented in the GAMESS package⁵⁸ and Exner and Mezey's field-adapted adjustable density matrix approach.^{59–63}

Molecular energy-based fragmentation utilizing the methods of Collins or of our group offers an intriguing possibility when applied to molecules like proteins. Proteins, being composed of around 20 amino acids, when fragmented into their constituent molecules produces a finite number of possible fragment molecules. This particular flavor of fragmentation breaks large molecules up based solely on the primary sequence. Thus if a database of precomputed highly accurate ab initio potential energy surfaces were available for these fragment molecules, then the “bonded” energy of proteins could be determined at comparatively negligible computational expense. The generation of the required highly dimensional potential energy surfaces is feasible because the fragment molecule sizes involved are relatively small. Of course, nonbonded interactions play a crucial role in such systems, but such interactions may be taken into account via electrostatics (through the utilization of distributed multipoles), induction (through distributed or central polarizabilities), and dispersion (through the real dynamic polarizabilities at imaginary frequencies), if fragment molecules are separated far enough from each other. All of these quantities can be determined from first principle calculations without the need of any parametrization and are solely a property of the monomer fragments.

Therefore apart from the electronic energy, distributed electrostatic multipoles, polarizabilities, and the dynamic polarizabilities at imaginary frequencies can also form part of the precomputed database of fragment molecules. Close-contact nonbonded interactions, most importantly H-bonds, could not be dealt with in such a manner. However, for these types of interactions, precomputed potential energy surfaces can also be generated because there is a finite number of possible close-contact interactions. Similarly, the solvent must be taken into account in such systems, and this may be done either implicitly or explicitly. For the latter, precomputed potential energy surfaces for water–water and small water clusters as well as water–protein fragment molecules can be determined for the close-contact interactions. Therefore, in principle at least, it may be possible to produce highly accurate second generation modeling software capable of dealing with protein–substrate interactions in the presence of solvent from first principles, but utilizing the fragmentation approximation, to replace the current empirical force field or QM/MM methods. Such a goal, if even possible, lies

far from the present, but a necessary step along the way is to show that highly dimensional potential energy surfaces can be constructed for possible fragment molecules. Furthermore, it is also necessary to show that distributed multipoles can be interpolated well enough to predict accurate electrostatics within the relevant configuration spaces of the fragments.

2. METHODOLOGY

2.1. Theory. *2.1.1. Distributed Multipoles.* In this work we utilized Stone's distributed multipole analysis,^{1,2} which has been well described elsewhere,^{2,64} so only a brief account will be provided here. The electrostatic potential at some location \mathbf{r} is given by the well-known expression:

$$V(\mathbf{r}) = \sum_{k=1}^{N_{\text{nuclei}}} \frac{Z_k}{|\mathbf{r} - \mathbf{r}_k|} - \int \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}' \quad (1)$$

where $\rho(\mathbf{r}')$ is the charge density which may incorporate the effects of electron correlation. The charge density takes the form

$$\rho(\mathbf{r}') = \sum_{st} P_{st} \chi_s(\mathbf{r}' - \mathbf{p}_s) \chi_t(\mathbf{r}' - \mathbf{p}_t) \quad (2)$$

where P_{st} is an element of the density matrix, and $\chi_s(\mathbf{r}' - \mathbf{p}_s)$ is a basis function centered at \mathbf{p}_s . The contribution made to the charge density from a product of Gaussian basis functions can be represented as a linear combination of multipoles, centered at \mathbf{p} , from rank zero to rank $l + l'$, where l and l' are the ranks of the original basis functions ($l = 0$, s -function; $l = 1$, p -function, etc.). The location of \mathbf{p} lies on a line in between the centers s and t and depends on the exponents of the basis functions. Thus the charge density resulting from the overlap of, say, two s -functions can be represented as a single monopole or the overlap of a d - and p -function as a monopole, dipole, quadrupole, and octapole only. However, the locations of each of these sets of multipoles will be different for each pair of basis functions. Each set of multipoles can have their origin shifted to one of the centers s or t or some other convenient center. The price paid for shifting the origin is the generation of an infinite number of multipoles at the new origin. The convergence of this infinite series of multipoles at the new center depends strongly on the distance shifted. How this shift is affected is ad hoc, but the smaller the shift, the more rapid the convergence of the multipole series. Thus there is a strong argument to make this shift to the nearest possible center, which may not be either site s or t .

In this way a set of multipoles can be generated at various locations within the molecule. The total molecular dipole of the molecule is exactly reproduced by the collection of distributed monopoles and dipoles – the molecular quadrupole by the collection of distributed monopoles, dipoles, and quadrupoles, etc. Furthermore the potential, electric field, electric field gradient, etc. of the molecule is very accurately reproduced at locations further away from the molecule than the van der Waals surface using a sufficient number of centers and multipoles.

The deficiencies in the above treatment lay in determining the multipoles piecemeal from the individual products of basis functions rather than from the molecular charge density in physical space and was most pronounced when the basis included

diffuse functions. The deficiency was not one of accuracy, but rather the distributed multipoles could vary widely and unpredictably by improving the basis set used to describe the molecule even though the electrostatic potential may have changed little.² This deficiency was alleviated in 2005 by modifying the distributed multipole analysis to allow for a numerical integration of the charge density due to diffuse, or extended, basis functions around atomic sites² to determine their contribution to the multipole moments at those sites.

2.1.2. Modified Shepard Interpolation. The modified Shepard interpolation^{15–18} has been used to accurately represent ab initio potential energy surfaces by the Collins group and others for both classical and quantum reaction dynamics (e.g., refs 16 and 65–67). It has also been utilized in stationary state problems (e.g., refs 68 and 69). In this work we employed the technique in order to interpolate both the energy and the distributed multipoles. Note that the interpolated surface always passes exactly through all data used in the interpolation.

The interpolation of some quantity, $X(\mathbf{Z})$, which is a function of the internal coordinates, \mathbf{Z} , proceeds by first expanding $X(\mathbf{Z})$ as a Taylor series about some specific location \mathbf{Z}_0 :

$$X(\mathbf{Z}) = X_{\mathbf{Z}_0} + \left. \frac{\partial X}{\partial \mathbf{Z}} \right|_{\mathbf{Z}_0} \cdot (\mathbf{Z} - \mathbf{Z}_0) + \frac{1}{2} (\mathbf{Z} - \mathbf{Z}_0)^T \left. \frac{\partial^2 X}{\partial \mathbf{Z}^2} \right|_{\mathbf{Z}_0} (\mathbf{Z} - \mathbf{Z}_0) + \dots \quad (3)$$

The Taylor series is then truncated at some order, and in this work it was truncated after the second order. Thus the estimate of X is only expected to be accurate in the vicinity around \mathbf{Z}_0 . Indeed, for some applications, like the crystal structure prediction of small molecules,⁷⁰ a single Taylor series expansion has been utilized, where X represents the electronic energy. However, the estimate of the quantity X can be improved through the addition of further Taylor series expanded about different locations. If we include all of the Taylor series estimates of X at \mathbf{Z} , then we may write

$$\chi(\mathbf{Z}) = \sum_{i=1}^N w_i(\mathbf{Z}) T_i(\mathbf{Z}) \quad (4)$$

where $T_i(\mathbf{Z})$ is a truncated Taylor series expanded about location \mathbf{Z}_i , and $w_i(\mathbf{Z})$ is the normalized weight given to Taylor series i , which depends upon the location \mathbf{Z} . N is the total number of Taylor series which constitute the interpolation data set. In our work, the simple “one-part” weight function¹⁵ (see eq 6) was used to add the first 40 data points for both Gly and Ala. After that, the more flexible “two-part” weight function (see eq 7) together with the confidence radius¹⁷ (see eq 8) were employed to improve the accuracy of the data sets:

$$w_i(\mathbf{Z}) = \frac{v_i(\mathbf{Z})}{\sum_{j=1}^N v_j(\mathbf{Z})} \quad (5)$$

where

$$v_i(\mathbf{Z}) = |\mathbf{Z} - \mathbf{Z}_i|^{-2p} \quad (6)$$

for the “one-part” weight function and

$$v_i(\mathbf{Z}) = \left\{ \left[\frac{|\mathbf{Z} - \mathbf{Z}_i|^{2q}}{\text{crad}_i} \right] + \left[\frac{|\mathbf{Z} - \mathbf{Z}_i|^{2p}}{\text{crad}_i} \right] \right\}^{-1} \quad (7)$$

for the “two-part” weight function. In this expression crad_i is given by

$$\text{crad}_i^{-6} = \frac{1}{N_{\text{neigh}}} \sum_{k=1}^{N_{\text{neigh}}} \frac{[X(\mathbf{Z}_k) - T(\mathbf{Z}_k)]^2}{E_{\text{tol}}^2 |\mathbf{Z}_k - \mathbf{Z}|^6} \quad (8)$$

N_{neigh} is the number of nearby configurations, $2p = 16$ and $q = 2$ for Gly and Ala, and E_{tol} was set to 0.2 m-Eh. Note that $2p$ must be greater than the number of degrees of freedom in order for the potential to converge using both one- and two-part weight functions. The number of degrees of freedom of Gly and Ala were 12 and 15, respectively (see the Fragment Structures section for an explanation of these numbers). For convenience and accuracy, $q = 2$ was taken.¹⁸ The confidence radius crad_i represents the distance away from Taylor series i in which the average error increases to the value of the error tolerance; crad_i is discussed in ref 17.

2.1.3. Axis Systems and “Axis-Free” Multipoles. The Shepard interpolation, as described above, is utilized to interpolate scalar quantities. In this work we have directly applied it to the interpolation of the total electronic energy and the distributed multipoles. The individual components of the distributed dipoles and quadrupoles might also be Shepard interpolated. That is, treated as though they were scalar quantities, but as we shall see later, imposing an axis system on the molecule and then interpolating the individual components leads to difficulties that can be avoided. It should be noted that multipoles are tensors, and thus they transform according to the tensor transformation law. This transformation affects the values of the components and therefore the interpolation.

The orientation of a molecule with respect to some lab-based Cartesian frame is defined in terms of the three Euler angles that rotate the lab-based Cartesian frame into the molecule-based Cartesian frame. For rigid molecules, i.e., those with no change in internal coordinates, when the molecule rotates, i.e., changes its orientation, the molecule-based Cartesian framework rotates with it. However, if the molecule can distort via some change in the internal coordinates, then the molecule-based Cartesian framework will generally move both in origin and orientation with respect to the lab-based framework. Thus, the three Euler angles depend upon the internal coordinates. This phenomenon is well-known in spectroscopy and is the origin of rovibrational coupling. It is therefore best to refer all components of the multipole moments to a molecule-based frame. The advantage of referring multipole components to this frame is that these components do not depend upon the Euler angles.

However, the disadvantage of referring the multipole components to a molecule-based frame is that if a poor choice is made for the frame, the components of the multipoles may vary significantly at atomic sites located far from the place in the molecule where a distortion has occurred. This variation in the components would be solely due to the changing of the molecule-based frame rather than any change in electron density at the remote atomic sites. One way to significantly alleviate this effect would be to define a local atomic Cartesian frame at each atomic site, say by using the atoms adjacent to the site in question. Such axis systems have been suggested and utilized before.¹²

Alternatively the molecule-based Cartesian frame can be disposed of entirely, and the distributed dipoles and quadrupoles expressed in terms of the internal coordinates of the molecule. It is this approach that we have adopted here. In order to achieve

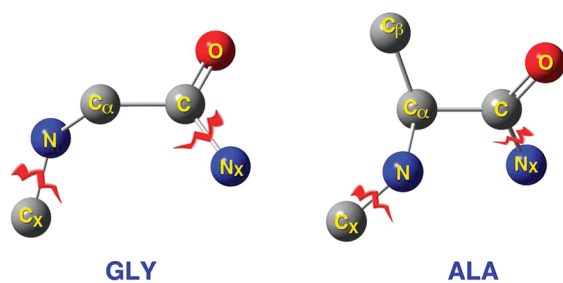


Figure 1. Gly and Ala residues extracted from a .pdb file. Also illustrated are the adjacent carbon and nitrogen atoms (labeled C_X and N_X) and the bonds to be cut and capped with hydrogen atoms.

this transformation we envisaged an atomic dipole as a vector which locates the position of a fictitious atom attached to the atomic site where the dipole originated. As such, the magnitude of the dipole represented a “bond length” to the fictitious atom. The direction of the dipole was expressed as a bond angle and a dihedral angle, which was made between the fictitious atom, the atomic site, and one (for angle) or two (for dihedral angle) nearby atoms. The Shepard interpolation then involved interpolating these three strictly scalar quantities in terms of the actual internal coordinates of the real atoms in the molecule.

The distributed quadrupoles were treated similarly but in a slightly more complicated manner. The quadrupole tensor and atomic coordinates in the standard orientation defined in the Gaussian software⁷¹ (utilized in this work) were first transformed into another molecule fixed axis system defined using the three atoms, N, C_ω and C (see Figure 1). This was necessary due to axis switching that arbitrarily occurred during numerical displacements in internal coordinates while calculating the derivatives and between different conformations.

The Cartesian quadrupole tensor in the new axis system was then diagonalized to obtain the principle quadrupole axes. We then considered each of the axes as a vector, and as such, the treatment was similar to that of the dipoles. The first two eigenvalues, with the last eigenvalue being equal to minus their sum, were readily treated as scalar quantities and interpolated as usual. The corresponding first two eigenvectors were described by placing two fictitious atoms at the end points of these vectors from the atomic site where the quadrupole originated. After placing these two fictitious atoms, three independent internal coordinates were defined to locate them: a single “bond angle” and two “dihedral angles”. The bond angle was defined to be between the first fictitious atoms, the atom to which it was “bonded”, and some adjacent atom in the molecule. Since the two eigenvectors are orthogonal it was unnecessary to define a second “bond angle”. The two “dihedral angles” were made to adjacent atoms within the molecule. These three coordinates completely specified the directions of the principle quadrupole axes. The Shepard interpolation then involved interpolating the five quantities: two eigenvalues, a “bond angle”, and two “dihedral angles”, in terms of the actual internal coordinates of the real atoms in the molecule.

Nevertheless, it should be noted that the eigenvectors are fundamentally bidirectional. As a consequence, there are two possible positions for each of the fictitious atoms. Hence the axes may arbitrarily switch between different conformations. For similar geometries this is readily detected and rectified. Unfortunately, for geometries located far from others in configuration

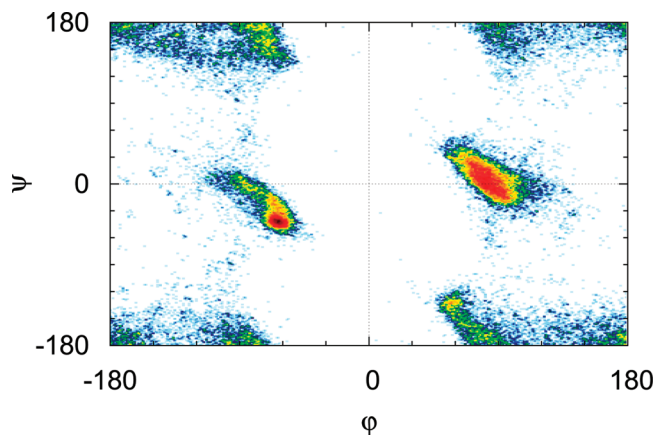


Figure 2. A 2-dimensional projection (φ and ψ) of the 12-dimensional data set extracted from the .pdb files for Gly. Red indicates high data density, while blue is low data density.

space this becomes problematic. A solution was found by erecting a network of nearest neighbors between all conformations. If a complete weighted graph is associated with the interpolation data set, a vertex being a conformation and the weight of an edge being the Euclidean distance in internal coordinates between them, then the desired network is a minimum spanning tree. In our work, Prim’s algorithm^{72,73} was used to determine one of the minimum spanning trees. The first vertex was defined as an end point of the first edge added during the construction of this tree. The orientation of the eigenvectors chosen for this starting vertex was then used as a point of reference to assign the orientation of the eigenvectors for subsequent geometries. This pretreatment assures consistency among the eigenvectors prior to interpolation, provided that the data points are not sparsely distributed. Therefore, a sufficient data density was necessary.

For both the dipole and quadrupole, interpolation of the dihedral angles to the fictitious atoms requires an additional comment. Because of the periodic nature of an angle, care must be taken when two or more Taylor series estimated a dihedral to be approximately $\pm \pi$. While the estimated angles may be nearly the same, the numerical values from individual Taylor series may have differed in sign so that application of eq 4 resulted in substantially different interpolated values for the dihedral.

2.2. Approach. **2.2.1. Fragment Structures.** The Research Collaboratory for Structural Bioinformatics protein data bank⁷⁴ was searched on October 12, 2009 for X-ray crystallographic structures of proteins with a resolution in the range of 0–1.3 Å. This resulted in a total of 1745 PDB files (listed in the Supporting Information). All structures were then searched within each file to obtain every occurrence of amino acid bracketed Gly and Ala. Only complete structures were accepted, and if relevant, those with an “alternative location indicator” of type “A”. Identical or near identical structures were removed. Our definition of “near identical” structures were those geometries \mathbf{R}_i , where \mathbf{R}_i is a vector of interatomic distances between heavy atoms (see Figure 1), which were closer than $|\mathbf{R}_i - \mathbf{R}_j| < 10^{-2}$ Å. Furthermore, “unlikely” geometries were also removed. Here “unlikely” means those geometries so distorted that it would seem to us unlikely that they would appear in any real protein. The criterion used was energy based (at the HF/6-31G level) and were all those structures higher in energy than 35 m-Eh for Gly and 44

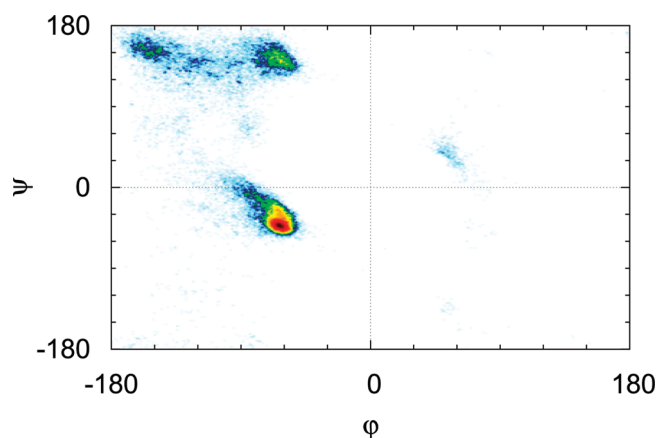


Figure 3. A 2-dimensional projection (ϕ and ψ) of the 15-dimensional data set extracted from the .pdb files for Ala. Red indicates high data density while blue is low data density.

m-Eh for Ala above the lowest energy structure encountered in the data set. These conditions excluded 37 and 18 geometries for Gly and Ala, respectively. After the above preparation of the data sets, the final number of geometries accepted for Gly and Ala were 41 544 and 42 849, respectively, with standard deviations in their energies of 3.42 and 3.53 m-Eh, respectively. Figures 2 and 3 show the 2-dimensional projections (ϕ and ψ) of the 12- and 15-dimensional data sets extracted for Gly and Ala, respectively. It is gratifying to note that these figures bear a striking resemblance to “textbook” Ramachandran plots for these residues.

Due to the nature of X-ray crystallography with its propensity to place hydrogen atoms too close to heavy atoms, all H atoms were removed from Gly and Ala residues on those few occasions when they were available. Because the intention of these extracted geometries is to form part of a fragment database, the carbon atom (C_x) adjacent to the nitrogen (N) and the nitrogen atom (N_x) adjacent to the carbonyl carbon (C) were also extracted because it is these two bonds (C_x-N and N_x-C) that will be broken and “capped” with hydrogen atoms to form the fragment residue (see Figure 1), an amino aldehyde.

Addition of H atoms to the heavy atoms was precisely and uniquely determined from the coordinates of the heavy atoms. Addition of H atoms to the Gly residue proceeded as follows. The capping H atom attached to N (see Figure 1) was located along the $N-C_x$ bond at a distance of $r_{NH}^0(r_{NCX}/r_{NC_x}^0)$ from N, where $r_{NC_x}^0 = 1.32 \text{ \AA} = r_{CN_x}^0$ and $r_{NH}^0 = 0.99 \text{ \AA}$. Likewise for the capping H attached to C, except $r_{CH}^0 = 1.07 \text{ \AA}$ was used. The remaining H atom attached to N was placed along the negative of the average of the unit vectors \hat{e}_{NC_x} and \hat{e}_{NC_α} with a bond length of r_{NH}^0 . The two H atoms attached to the C_α were placed a distance r_{CH}^0 along unit vectors whose end points were equally distant from each other and the end points of the unit vectors of $\hat{e}_{C_\alpha N}$ and $\hat{e}_{C_\alpha C}$.

Addition of H atoms to the Ala residue differed from Gly at the C_α only. For Ala there is an additional C atom bonded to C_α namely C_β . The single H atom attached to C_α was located r_{CH}^0 in the negative direction of the average of the three unit vectors, $\hat{e}_{C_\alpha N}$, $\hat{e}_{C_\alpha C}$, and $\hat{e}_{C_\alpha C_\beta}$. Having placed the H_α atom, the three H atoms of the C_β can be uniquely located. The first H atom, $H_{\beta 1}$, was placed anti to H_α using the tetrahedral angle for $\angle C_\alpha C_\beta H_{\beta 1}$ and at a distance of r_{CH}^0 from C_β . The remaining two H atoms, $H_{\beta 2}$ and $H_{\beta 3}$, were located such that the methyl group possessed C_{3v} symmetry with the C_3 axis lying along $C_\beta-C_\alpha$.

Note that the numbers of degrees of freedom available to Gly and Ala were reduced because the positions of the H atoms were entirely dependent on the heavy atom coordinates. Thus the numbers of degrees of freedom for Gly and Ala were 12 and 15, respectively. In generating the internal coordinates used in the Taylor series expansion, bond lengths, angles, and dihedrals were all referred to heavy atoms.

2.2.2. Fragment Energies, Distributed Multipoles, and Electrostatic Potentials. To facilitate and assess our method for interpolating energies and distributed multipoles, the energies and multipoles of the entire Gly and Ala data sets were determined at the HF/6-31G level using the Gaussian 03 suite of programs.⁷¹ Note that Shepard interpolation is entirely independent of level of theory, therefore it is only necessary for us to investigate our approach at the above crude level of theory in order to establish its validity and accuracy. Of course in the generation of an actual working database a much higher level of theory, including post-Hartree–Fock effects, would be utilized.

While we may directly compare interpolated energies with the ab initio energies to assess accuracy, comparing interpolated multipoles to those derived directly from the ab initio calculations does not provide much insight into the actual error generated, say, in the potential at meaningful locations in the vicinity of the molecule. In order to assess the success of interpolating the multipoles we have computed the electrostatic potential at the solvent accessible surface using: (a) the interpolated multipoles and (b) those derived from a distributed multipole analysis as well as (c) directly from the electronic wave function. Comparison of (b) with (c) provides us with an indication of the errors associated with using distributed multipoles for predicting the electrostatic potential. Comparison of (a) with (b) provides us with an indication of the errors associated with the interpolation.

For each configuration in the data set, points were placed on the solvent accessible surface at a density of approximately 1 point/ \AA^2 . The solvent accessible surface was located using a probe radius of 1.4 \AA and the Bondi van der Waals radii of 1.20, 1.70, 1.55, and 1.52 \AA for H, C, N, and O respectively.⁷⁵ The algorithm used to locate the solvent accessible surface was essentially that found in Appendix II of ref 76 except applied to the solvent accessible surface rather than the molecular surface. The average surface areas of Gly and Ala were 210 and 236 \AA^2 , respectively.

A summary of the error in the computed potential at each point on the solvent accessible surface can be expressed as an root mean square (RMS) and an RRMS⁷⁷ for an individual fragment or over the entire data set. These quantities are defined below for the entire data set. For an individual fragment, i , the sum over i is excluded.

$$V_{\text{RMS}} = \left[\frac{1}{M} \sum_{i=1}^{N_{\text{frag}}} \sum_{j=1}^{N_i} (V_{i,j} - v_{i,j})^2 \right]^{1/2} \quad (9)$$

$$V_{\text{RRMS}} = \left[\frac{\sum_{i=1}^{N_{\text{frag}}} \sum_{j=1}^{N_i} (V_{i,j} - v_{i,j})^2}{\sum_{i=1}^{N_{\text{frag}}} \sum_{j=1}^{N_i} V_{i,j}^2} \right]^{1/2} \quad (10)$$

Here $V_{i,j}$ is the ab initio potential (or distributed multipoles potential) at point j in fragment i , and $v_{i,j}$ is the computed potential

for the same fragment and point using distributed multipoles (or interpolated distributed multipoles). M is the total number of points in the double summation, i.e., $M = \sum_i^{N_{\text{frag}}} N_i$, and N_i is the number of points on the solvent accessible surface for fragment i .

2.2.3. Importance Sampling. Having selected the first geometry about which to expand a Taylor series, addition of further Taylor series proceeded as follows. The most efficient approach would seem to be one which adds a Taylor series that maximally reduces the interpolation errors. We define the energy interpolation error as an RMS of the residuals associated with the energy. Thus

$$E_{\text{RMS}} = \left[\frac{1}{M} \sum_{i=1}^M \gamma_i^2 \right]^{1/2} \quad (11)$$

where $\gamma_i = E_i - \varepsilon_i$, and M is the number sampled points, which for Gly is 41 544. Addition of a Taylor series at geometry $\mathbf{Z}_i = \mathbf{Z}_a$ will at least eliminate all error associated with point i . If this Taylor series is in the near vicinity of many other similar geometries with significant error associated with them, then by adding this Taylor series to the interpolation data set we can also expect to reduce the interpolation error associated with the nearby geometries as well. Thus our goal is to select a geometry, about which we will expand a new Taylor series, that has significant error associated with it as well as many other neighboring geometries.

To proceed we note that the energy error in our truncated second-order Taylor series expanded about point a is

$$E_{\mathbf{Z}} - T_{\mathbf{Z}_a}(\mathbf{Z}) = \mathcal{O}(d_{\mathbf{Z}_a}(\mathbf{Z})^3) \approx f(d_{\mathbf{Z}_a}(\mathbf{Z})^3) \quad (12)$$

where $d_{\mathbf{Z}_a}(\mathbf{Z}) = |\mathbf{Z} - \mathbf{Z}_a|$ in the vicinity of \mathbf{Z}_a . We now imagine we have added a new Taylor series to our interpolation data set at configuration $\mathbf{Z} - \mathbf{Z}_a$ and compute our new RMS energy error, E'_{RMS} which can readily be shown to be

$$M \times E'_{\text{RMS}}{}^2 = \sum_{i=1}^M \left\{ \frac{s(i)}{s(i) + v_a(i)} \gamma_i + w_a(i) f(d_{\mathbf{Z}_a}(\mathbf{Z})^3) \right\}^2 \quad (13)$$

where $s(i) = \sum_{k=1}^N v_k(i)$, and N is the cardinality of the interpolation data set. If we imagine our new Taylor series to be particularly accurate, then we can set $f(d_{\mathbf{Z}_a}(\mathbf{Z})^3) \approx 0$, and we arrive at the expression used to select a new geometry about which we will expand a Taylor series:

$$t_a = \sum_{i=1}^M \left(\frac{s(i)}{s(i) + v_a(i)} \right)^2 \gamma_i^2 \quad (14)$$

Using the above expression we compute a t_a for each value of i , i.e., there will be a total of M different values of t_a . The best point to choose to perform a new Taylor series expansion is about that point corresponding to the smallest value of t_a . Such a location would be one where there is a large number of similar structures each possessing a relatively large value of γ_i^2 . In addition, once the point is determined that leads to the smallest value of t_a , the geometry could be further refined by minimizing t_a with respect to \mathbf{Z}_a . The geometry that minimized t_a would be chosen as the next point to add to the interpolation data set. By choosing such a point to add to the interpolation data set, we expect the greatest possible reduction in the RMS error.

Nevertheless, as the new Taylor series was assumed to be accurate, the sampling can become relatively inefficient later on

in the growing process. This is because already sampled points may lie in regions of high data density, but the sampled point has nonzero error associated with it. In this case the smallest value of t_a is obtained by replacing the already sampled point with the assumed zero error Taylor series. That is, the point selected to be added to the interpolation data set already exists in the data set. This problem was resolved by introducing an expression for $f(d_{\mathbf{Z}_a}(\mathbf{Z})^3)$ into the formula for t_a once the cardinality of the interpolation data set was large enough.

$$f(d_{\mathbf{Z}_a}(\mathbf{Z})^3) \approx E_{\text{tol}} \left(\frac{|\mathbf{Z}_a - \mathbf{Z}_i|}{\text{crad}_a} \right)^3 \quad (15)$$

where crad_a is the confidence radius at geometry \mathbf{Z}_a , which was assumed to be equal to $\min\{\text{crad}_i\}$, where i runs over the interpolation data set. The nonzero term $f(d_{\mathbf{Z}_a}(\mathbf{Z})^3)$ offers a means to incorporate into the t_a formula an approximate level of reliability of our existing Taylor series.

$$t_a = \sum_{i=1}^{N_{\text{neigh}}} \left[\frac{s(i)}{s(i) + v_a(i)} \gamma_i + w_a(i) E_{\text{tol}} \left(\frac{|\mathbf{Z}_a - \mathbf{Z}_i|}{\text{crad}_a} \right)^3 \right]^2 \quad (16)$$

and N_{neigh} is the number of neighboring configurations, here set to 1000. In our work, the first 80 data points for both Gly and Ala were added using eq 14; the rest were added using eq 16.

The above-described “ t_a method” of selecting a data point to add to the interpolation data set warrants further comment. It would seem that such a method of importance sampling a potential energy surface has little utility outside the present application. This is because the above expression requires the actual energy errors at all sampled geometries, while in general, such information is not available. However, related information is available in the form of an energy variance associated with each sampled point and has been described elsewhere as the “RMS method”. The expression that provides the variance in the predicted energy at a given location is⁷⁸

$$\sigma_E^2(\mathbf{Z}) = \sum_{j=1}^N w_j(\mathbf{Z}) [\varepsilon_{\mathbf{Z}} - T_j(\mathbf{Z})]^2 \quad (17)$$

Thus locations where the energy is predicted to be very different values by Taylor expansions possessing high weights are locations of high energy uncertainty. By substitution of eq 17 for γ_i^2 in t_a , we are able to importance sample in regions of configuration space that contributes most to the uncertainty of the interpolated potential energy surface. That is

$$\tau_a = \sum_{i=1}^M \left(\frac{s(i)}{s(i) + v_a(i)} \right)^2 \sum_{j=1}^N w_j(i) [\varepsilon_{\mathbf{Z}} - T_j(\mathbf{Z})]^2 \quad (18)$$

By selecting the point, i , that produces the smallest value of τ_a then minimizing with respect to \mathbf{Z}_a , we expect to obtain the best reduction in overall uncertainty in the interpolated potential energy surface.

It is of note that this approach may be superior to the previous RMS method of importance sampling a potential energy surface in reaction dynamics⁷⁹ and stationary state calculations.⁶⁸ This is because it will select a configuration associated with significant uncertainty in the interpolated surface at locations of high configuration density encountered while sampling. As such the above τ_a sampling method incorporates both the RMS- and h -weight⁸⁰ sampling methods within a single method and does not require any constraints in the minimization carried out in ref 79,

so this should improve the efficiency of sampling and thus reduce computational expense. We are planning to further investigate this method in a later publication.

3. RESULTS AND DISCUSSION

3.1. Interpolation Data Sets. The first point added to each of the interpolation data sets was the lowest energy configuration from the corresponding sample sets. The next 39 points were added using the one-part weight function and the t_a sampling method on the electronic energies. Continuing with t_a sampling on the electronic energy, the next 40 points utilized the two-part

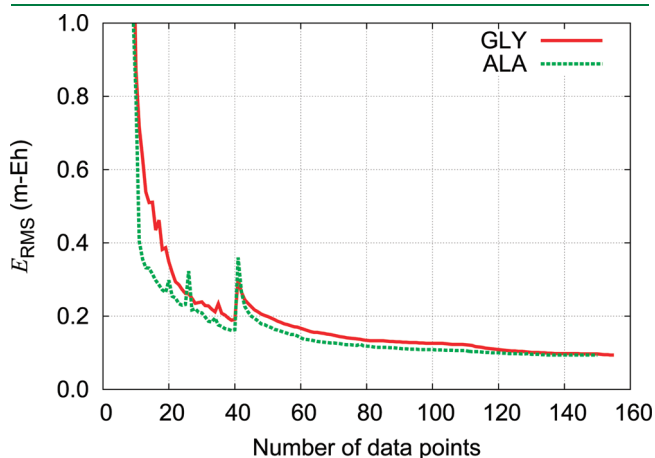


Figure 4. E_{RMS} for Gly and Ala as a function of the number of data points.

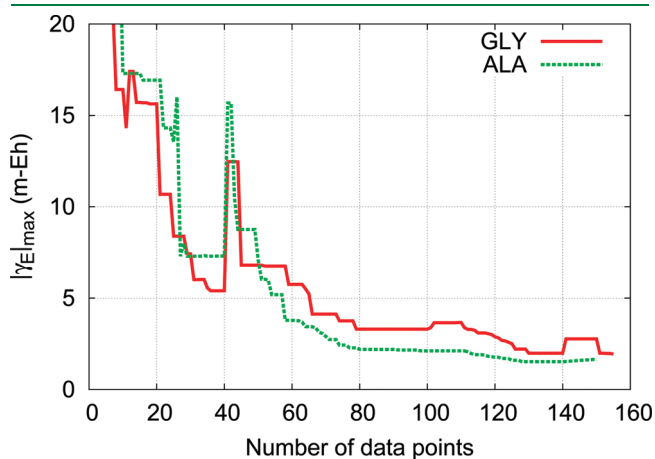


Figure 5. $|\gamma_E|_{\text{max}}$ for Gly and Ala as a function of the number of data points.

weight function. Following this, 30 points utilized t_a sampling in the electronic potential at the solvent accessible surface, then 20 points were added by sampling on the energy, and finally 20 further points were added by sampling on the potential. Additionally for Gly a further five points were sampled using the t_a method on the electronic energy. Thus the final cardinality of the interpolation data sets was 150 for Ala and 155 for Gly. Figures 4 and 5 show the RMS error in the energy, E_{RMS} , and the maximum absolute error in the energy, $|\gamma_E|_{\text{max}}$, respectively, as a function of the cardinality of the interpolation data sets.

As expected, the t_a method smoothly reduces the E_{RMS} and $|\gamma_E|_{\text{max}}$. A peak was observed in both plots at 40 data points as a result of switching from one- to two-part weight function. The RMS error as well as the maximum absolute error in the energy using our final interpolation data sets were evaluated to be respectively 0.094 and 1.942 m-Eh for Gly and 0.094 and 1.654 m-Eh for Ala. Thus using less than 0.4% of the sample set, the modified Shepard interpolation is capable of reproducing the vast majority of electronic energies to better than 0.1 m-Eh with no sampled configuration possesses an electronic energy error greater than 2 m-Eh. This level of accuracy should be sufficient for even the most demanding applications.

3.2. The Electrostatic Potential. The accuracy in reproducing the ab initio electrostatic potential using the modified Shepard interpolation depends upon several factors. First, since we are utilizing distributed multipoles to compute the electrostatic potential, we need to evaluate their accuracy at doing so for a given multipole rank and site selection. Second, the errors associated with interpolating the multipoles also impacts on how well the ab initio potential can be reproduced. We require that this second contribution to the errors in the potential at the solvent accessible surface to be minimized or even negligible in comparison to the errors associated with the first.

Previously we concluded from the results of several test molecules that rank two multipoles were sufficient to obtain errors less than or about equal to 1 m-Eh in the predicted electrostatic potential around the solvent accessible surface.⁵⁴ We wish to verify that this is the case for the two selected sample sets here. As such, distributed multipoles up to rank five were computed from the HF/6-31G wave function using the GDMA2 program² for all geometries in each sample set. As described in the Approach Section, the potential was computed at points on the solvent accessible surface using a density of about 1/Å, and the V_{RMS} and V_{RRMS} were evaluated. For Gly and Ala this amounted to about 9×10^6 and 10×10^6 points, respectively. A summary of the results is provided in Table 1.

Not surprisingly, Table 1 shows that the distributed multipoles are capable of producing near exact agreement with the potential

Table 1. V_{RMS} and V_{RRMS} Error at the Solvent Accessible Surface in the Electrostatic Potential between That Computed with Distributed Multipoles to the Rank Indicated and the ab Initio Potential

rank	Gly all atoms		Gly heavy and cap hydrogens		Ala all atoms		Ala heavy and cap hydrogens	
	V_{RMS}	V_{RRMS} (%)	V_{RMS}	V_{RRMS} (%)	V_{RMS}	V_{RRMS} (%)	V_{RMS}	V_{RRMS} (%)
0	4.12	27.45	21.88	145.49	4.11	29.86	22.29	162.13
1	3.91	26.02	4.41	29.39	3.61	26.26	3.98	28.95
2	0.64	4.27	0.89	5.94	0.64	4.64	0.85	6.21
3	0.09	0.59	0.23	1.51	0.11	0.77	0.27	1.94
4	0.04	0.28	0.11	0.74	0.05	0.34	0.15	1.06
5	0.02	0.11	0.05	0.34	0.02	0.16	0.08	0.58

Table 2. V_{RMS} and V_{RRMS} Errors between the Electrostatic Potential at the Solvent Accessible Surface Computed with the Interpolated Distributed Multipoles and That Computed by the Exact Distributed Multipoles as well as ab Initio Potential to the Rank Indicated

rank	compared to Stone's exact DMs				compared to ab initio potential			
	Gly		Ala		Gly		Ala	
	V_{RMS}	V_{RRMS} (%)	V_{RMS}	V_{RRMS} (%)	V_{RMS}	V_{RRMS} (%)	V_{RMS}	V_{RRMS} (%)
0	0.044	0.23	0.050	0.24	21.881	145.79	22.289	162.12
1	0.085	0.57	0.085	0.61	4.413	29.40	3.982	28.96
2	0.157	1.04	0.241	1.75	0.904	6.02	0.895	6.51

Table 3. V_{RMS} Error at the Solvent Accessible Surface in the Electrostatic Potential for Each Atom between That Computed with Distributed Multipoles at Rank Two

Gly		Ala	
N	0.204	N	0.313
C_{α}	0.223	C_{α}	0.452
C	0.149	C	0.293
O	0.108	O	0.174
H_{CX}	0.160	C_{β}	0.225
$H_{\alpha 2}$	0.163	H_{CX}	0.223
$H_{\alpha 3}$	0.155	H_{α}	0.334
H	0.203	$H_{\beta 1}$	0.213
H_{NX}	0.132	$H_{\beta 2}$	0.207
		$H_{\beta 3}$	0.214
		H	0.284
		H_{NX}	0.243
V_{RMS} (m-au)	0.157		0.241

obtained from the electronic wave function at the solvent accessible surface. V_{RMS} errors as low as $20 \mu\text{-au}$ can be obtained with rank five multipoles located on all nuclei. It is noted that only a small reduction in accuracy is obtained if multipoles are centered on heavy and capping hydrogens, except in the case of distributed charges. However, in a situation where a database is to be used, say, to perform molecular dynamics or Monte Carlo simulations, it would seem that reducing the rank of the multipoles and the number of sites, while still provide adequate accuracy, would be best to select. It is evident from Table 1 that rank two with multipole sites placed on heavies and capping hydrogens represents a good trade-off, as the accuracy is still better than 1 m-au.

Next we consider the contribution of the errors directly associated with the modified Shepard interpolation. The potential on the solvent accessible surface was computed from the interpolated multipoles and then compared to both the ab initio potential and the potential computed from the exact multipoles at the same rank. V_{RMS} and V_{RRMS} were evaluated. Indeed, it is evident from Table 2 that the potential computed from the interpolated multipoles agrees well with that using the exact multipoles. Carefully comparing Tables 1 and 2 shows that our interpolated potential leads to negligible additional error other than that produced by using exact rank two multipoles placed on heavy atoms and capping hydrogens.

Finally we examine how well the electrostatic potential can be reproduced compared to Stone's exact distributed multipoles in

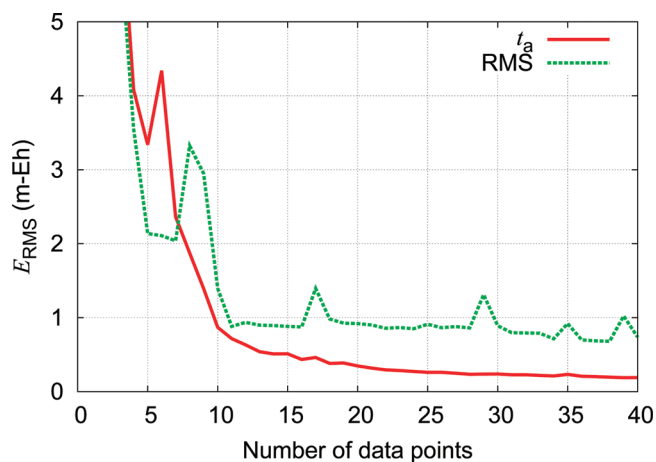


Figure 6. E_{RMS} using t_a and RMS methods as a function of the number of data points.

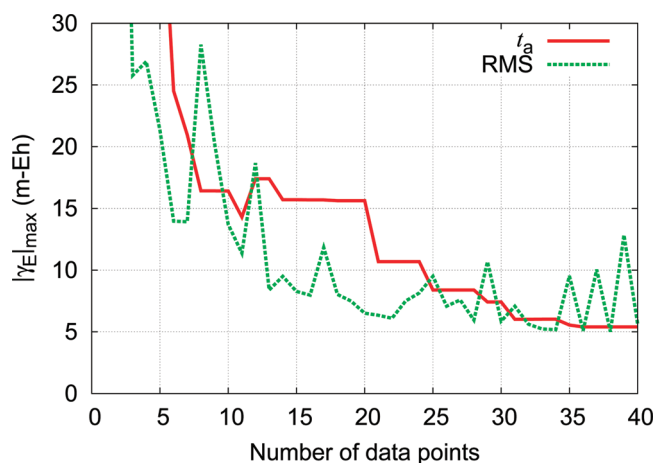


Figure 7. $|\gamma_E|_{\text{max}}$ using t_a and RMS methods as a function of the number of data points.

regions associated with each atom on the solvent accessible surface. The results are provided in Table 3. It is clear from this table that there are no regions around either molecule that are described particularly poorly with our method.

3.3. The Importance Sampling Method. To verify that our t_a method is superior to the previous RMS method, E_{RMS} and $|\gamma_E|_{\text{max}}$ were computed for the first 40 data points of the Gly potential energy surface using these two methods. The results are

illustrated in Figures 6 and 7. It is apparent that the t_a method achieves much lower E_{RMS} and comparable $|\gamma_E|_{\text{max}}$. Moreover, the fluctuations appearing in both figures for the RMS method are more frequent and pronounced, which implies more instability in the early stages of “growing” the potential energy surface. Worse is that for the same number of data points, beyond a very small number, the E_{RMS} using the RMS method is approximately four times greater than that obtained using the t_a method, implying greater computational expense in generating the potential energy surface.

4. CONCLUSION

We showed that for the 12- and 15-dimensional systems of the amino aldehydes, Gly and Ala, that the RMS energy error in electronic energies can be interpolated to better than 0.1 m-Eh for more than 41 000 different configurations encountered in protein X-ray structures. We also showed that distributed multipoles up to and including rank two can be interpolated very accurately so that negligible additional error is introduced into the calculation of electrostatic potential generated at the solvent accessible surface by the exact distributed multipoles. Rank two distributed multipoles lead to less than 1 m-au error in the potential at the solvent accessible surface. Considerable improvement in this accuracy was obtained by including rank three multipoles. The modified Shepard interpolation used in determining the interpolated energies and distributed multipoles required a small number of configurations selected using a newly described efficient sampling method, the “ t_a method”. This small number of points was selected from a set of over 41 000 different configurations encountered in protein X-ray data. Multipoles were also interpolated in an “axis-free” manner, which alleviated difficulties encountered in interpolating Cartesian components.

■ ASSOCIATED CONTENT

S **Supporting Information.** A complete listing of all the .pdb files used in this work can be found in Table S1. Also included in the Supporting Information are the Cartesian coordinates of the heavy atoms of all of the molecules included in the interpolation data sets for Gly (Table S2) and Ala (Table S3). This material is available free of charge via the Internet at <http://pubs.acs.org/>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: chmbrpa@nus.edu.sg.

■ ACKNOWLEDGMENT

The authors thank the National University of Singapore's support from the Academic Research Fund, grant number R-143-000-402-112. The authors also thank the Centre for Computational Science and Engineering for the use of their computers.

■ REFERENCES

- (1) Stone, A. J.; Alderton, M. *Mol. Phys.* **1985**, *56*, 1047–1064.
- (2) Stone, A. J. *J. Chem. Theo. Comput.* **2005**, *1*, 1128–1132.
- (3) Kosov, D. S.; Popelier, P. L. A. *J. Phys. Chem. A* **2000**, *104*, 7339–7345.
- (4) Whitehead, C. E.; Breneman, C. M.; Sukumar, N.; Ryan, M. D. *J. Comput. Chem.* **2003**, *24*, 512–529.

- (5) Sokalski, W. A.; Poirier, R. A. *Chem. Phys. Lett.* **1983**, *98*, 86–92.
- (6) Sokalski, W. A.; Sawaryn, A. *J. Chem. Phys.* **1987**, *87*, 526–534.
- (7) Koch, U.; Stone, A. J. *J. Chem. Soc. Faraday Trans.* **1996**, *92*, 1701–1708.
- (8) Brodersen, S.; Wilke, S.; Leusen, F. J. J.; Engel, G. *Phys. Chem. Chem. Phys.* **2003**, *5*, 4923–4931.
- (9) Koch, U.; Popelier, P. L. A.; Stone, A. J. *Chem. Phys. Lett.* **1995**, *238*, 253–260.
- (10) Karamertzanis, P. G.; Price, S. L. *J. Chem. Theo. Comput.* **2006**, *2*, 1184–1199.
- (11) Price, S. L.; Leslie, M.; Welch, G. W. A.; Habgood, M.; Price, L. S.; Karamertzanis, P. G.; Day, G. M. *Phys. Chem. Chem. Phys.* **2010**, *12*, 8478–8490.
- (12) Ren, P. Y.; Ponder, J. W. *J. Comput. Chem.* **2002**, *23*, 1497–1506.
- (13) Ren, P. Y.; Ponder, J. W. *J. Phys. Chem. B* **2003**, *107*, 5933–5947.
- (14) Liang, T.; Walsh, T. R. *Phys. Chem. Chem. Phys.* **2006**, *8*, 4410–4419.
- (15) Ischtwan, J.; Collins, M. A. *J. Chem. Phys.* **1994**, *100*, 8080–8088.
- (16) Collins, M. A. *Theo. Chem. Acc.* **2002**, *108*, 313–324.
- (17) Bettens, R. P. A.; Collins, M. A. *J. Chem. Phys.* **1999**, *111*, 816–826.
- (18) Thompson, K. C.; Jordan, M. J. T.; Collins, M. A. *J. Chem. Phys.* **1998**, *108*, 8302–8316.
- (19) Devereux, M.; Popelier, P. L. A.; McLay, I. M. *J. Comput. Chem.* **2009**, *30*, 1300–1318.
- (20) Rahalkar, A. P.; Ganesh, V.; Gadre, S. R. *J. Chem. Phys.* **2008**, *129*, 234101.
- (21) Ganesh, V.; Dongare, R. K.; Balanarayan, P.; Gadre, S. R. *J. Chem. Phys.* **2006**, *125*, 104109.
- (22) Gadre, S. R.; Shirsat, R. N.; Limaye, A. C. *J. Phys. Chem.* **1994**, *98*, 9165–9169.
- (23) Babu, K.; Gadre, S. R. *J. Comput. Chem.* **2003**, *24*, 484–495.
- (24) Babu, K.; Ganesh, V.; Gadre, S. R.; Ghermani, N. E. *Theo. Chem. Acc.* **2004**, *111*, 255–263.
- (25) Zhang, D. W.; Zhang, J. Z. H. *J. Chem. Phys.* **2003**, *119*, 3599–3605.
- (26) Mei, Y.; Ji, C.; Zhang, J. Z. H. *J. Chem. Phys.* **2006**, *125*, 094906.
- (27) Chen, X. H.; Zhang, J. Z. H. *J. Chem. Phys.* **2006**, *125*, 044903.
- (28) Chen, X. H.; Zhang, Y. K.; Zhang, J. Z. H. *J. Chem. Phys.* **2005**, *122*, 184105.
- (29) Zhang, D. W.; Zhang, J. Z. H. *Int. J. Quantum Chem.* **2005**, *103*, 246–257.
- (30) He, X.; Zhang, J. Z. H. *J. Chem. Phys.* **2005**, *122*.
- (31) Mei, Y.; Zhang, D. W.; Zhang, J. Z. H. *J. Phys. Chem. A* **2005**, *109*, 2–5.
- (32) Chen, X. H.; Zhang, J. Z. H. *J. Theo. Comput. Chem.* **2004**, *3*, 277–289.
- (33) Gao, A.; Zhang, D. W.; Zhang, J. Z. H.; Zhang, Y. K. *Chem. Phys. Lett.* **2004**, *394*, 293–297.
- (34) Xiang, Y.; Zhang, D. W.; Zhang, J. Z. H. *J. Comput. Chem.* **2004**, *25*, 1431–1437.
- (35) Chen, X. H.; Zhang, J. Z. H. *J. Chem. Phys.* **2004**, *120*, 11386–11391.
- (36) Zhang, D. W.; Xiang, Y.; Gao, A. M.; Zhang, J. Z. H. *J. Chem. Phys.* **2004**, *120*, 1145–1148.
- (37) Chen, X. H.; Zhang, D. W.; Zhang, J. Z. H. *J. Chem. Phys.* **2004**, *120*, 839–844.
- (38) Zhang, D. W.; Xiang, Y.; Zhang, J. Z. H. *J. Phys. Chem. B* **2003**, *107*, 12039–12041.
- (39) Zhang, D. W.; Chen, X. H.; Zhang, J. Z. H. *J. Comput. Chem.* **2003**, *24*, 1846–1852.
- (40) Li, S. H.; Li, W.; Fang, T. *J. Am. Chem. Soc.* **2005**, *127*, 7215–7226.
- (41) Dong, H.; Hua, S.; Li, S. *J. Phys. Chem. A* **2009**, *113*, 1335–1342.
- (42) Li, W.; Dong, H.; Li, S. Relative Energies of Proteins and Water Clusters Predicted with the Generalized Energy-Based Fragmentation

Approach. 12th European Workshop on Quantum Systems in Chemistry and Physics, London, England, August 30–September 5, 2007; Springer: 2008.

- (43) Hua, W.; Fang, T.; Li, W.; Yu, J.-G.; Li, S. *J. Phys. Chem. A* **2008**, *112*, 10864–10872.
- (44) Li, H.; Li, W.; Li, S.; Ma, J. *J. Phys. Chem. B* **2008**, *112*, 7061–7070.
- (45) Li, W.; Li, S.; Jiang, Y. *J. Phys. Chem. A* **2007**, *111*, 2193–2199.
- (46) Li, W.; Fang, T.; Li, S. *H. J. Chem. Phys.* **2006**, *124*, 154102.
- (47) Deev, V.; Collins, M. A. *J. Chem. Phys.* **2005**, *122*, 154102.
- (48) Collins, M. A.; Deev, V. A. *J. Chem. Phys.* **2006**, *125*, 104104.
- (49) Collins, M. A. *J. Chem. Phys.* **2007**, *127*, 024104.
- (50) Netzloff, H. M.; Collins, M. A. *J. Chem. Phys.* **2007**, *127*, 134113.
- (51) Bettens, R. P. A.; Lee, A. M. *J. Phys. Chem. A* **2006**, *110*, 8777–8785.
- (52) Lee, A. M.; Bettens, R. P. A. *J. Phys. Chem. A* **2007**, *111*, 5111–5115.
- (53) Bettens, R. P. A.; Lee, A. M. *Chem. Phys. Lett.* **2007**, *449*, 341–346.
- (54) Le, H.-A.; Lee, A. M.; Bettens, R. P. A. *J. Phys. Chem. A* **2009**, *113*, 10527–10533.
- (55) Yang, W. T.; Lee, T. S. *J. Chem. Phys.* **1995**, *103*, 5674–5678.
- (56) Nakano, T.; Kaminuma, T.; Sato, T.; Akiyama, Y.; Uebayasi, M.; Kitaura, K. *Chem. Phys. Lett.* **2000**, *318*, 614–618.
- (57) Fedorov, D. G.; Kitaura, K. *J. Phys. Chem. A* **2007**, *111*, 6904–6914.
- (58) Schmidt, M. W.; Baldrige, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. J.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347–1363.
- (59) Exner, T. E.; Mezey, P. G. *J. Phys. Chem. A* **2002**, *106*, 11791–11800.
- (60) Exner, T. E.; Mezey, P. G. *J. Comput. Chem.* **2003**, *24*, 1980–1986.
- (61) Exner, T. E.; Mezey, P. G. *J. Phys. Chem. A* **2004**, *108*, 4301–4309.
- (62) Exner, T. E.; Mezey, P. G. *Phys. Chem. Chem. Phys.* **2005**, *7*, 4061–4069.
- (63) Eckard, S.; Exner, T. E. *Int. J. Res. Phys. Chem. Chem. Phys.* **2006**, *220*, 927–944.
- (64) Stone, A. J. *The Theory of Intermolecular Forces*; Clarendon: Oxford, U.K., 2002.
- (65) Frankcombe, T. J.; Collins, M. A.; Worth, G. A. *Chem. Phys. Lett.* **2010**, *489*, 242–247.
- (66) Cao, J. W.; Zhang, Z. J.; Zhang, C. F.; Liu, K.; Wang, M. H.; Bian, W. S. *Proc. Nat. Acad. Sci. U.S.A.* **2009**, *106*, 13180–13185.
- (67) Wu, T.; Werner, H. J.; Manthe, U. *J. Chem. Phys.* **2006**, *124*, 164307.
- (68) Bettens, R. P. A. *J. Am. Chem. Soc.* **2003**, *125*, 584–587.
- (69) Yagi, K.; Oyanagi, C.; Taketsugu, T.; Hirao, K. *J. Chem. Phys.* **2003**, *118*, 1653–1660.
- (70) Kazantsev, A. V.; Karamertzanis, P. G.; Pantelides, C. C.; Adjiman, C. S. In *Molecular Systems Engineering*; Adjiman, C. S., Galindo, A., Eds.; Wiley: Weinheim, Germany, 2010; Vol. 6.
- (71) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*; Gaussian, Inc.: Wallingford, CT, 2004.
- (72) Prim, R. C. *Bell Syst. Tech. J.* **1957**, 1389–1401.
- (73) Jarník, V. *Acta Soc. Sci. Nat. Moraviae* **1930**, *6*, 57–63.
- (74) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235–242.
- (75) Bondi, A. *J. Phys. Chem.* **1964**, *68*, 441–451.
- (76) Connolly, M. L. *J. Appl. Crystallogr.* **1983**, *16*, 548–558.
- (77) Bayly, C.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269–10280.
- (78) Thompson, K. C.; Collins, M. A. *J. Chem. Soc. Faraday Trans.* **1997**, *93*, 871–878.
- (79) Moyano, G. E.; Collins, M. A. *J. Chem. Phys.* **2004**, *121*, 9769–9775.
- (80) Jordan, M. J. T.; Thompson, K. C.; Collins, M. A. *J. Chem. Phys.* **1995**, *102*, 5647–5657.

DFTB3: Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method (SCC-DFTB)

Michael Gaus,[†] Qiang Cui,[‡] and Marcus Elstner^{*,†}

[†]Institute of Physical Chemistry, Karlsruhe Institute of Technology, Kaiserstr. 12, 76131 Karlsruhe, Germany

[‡]Department of Chemistry and Theoretical Chemistry Institute, University of Wisconsin, Madison, 1101 University Avenue, Madison, Wisconsin 53706, United States

 Supporting Information

ABSTRACT: The self-consistent-charge density-functional tight-binding method (SCC-DFTB) is an approximate quantum chemical method derived from density functional theory (DFT) based on a second-order expansion of the DFT total energy around a reference density. In the present study, we combine earlier extensions and improve them consistently with, first, an improved Coulomb interaction between atomic partial charges and, second, the complete third-order expansion of the DFT total energy. These modifications lead us to the next generation of the DFTB methodology called DFTB3, which substantially improves the description of charged systems containing elements C, H, N, O, and P, especially regarding hydrogen binding energies and proton affinities. As a result, DFTB3 is particularly applicable to biomolecular systems. Remaining challenges and possible solutions are also briefly discussed.

1. INTRODUCTION

Recent years have shown that approximate quantum chemistry methods form an essential part in the repertoire of computational methods for an atomistic understanding of a broad range of physical, chemical, and biological problems. Besides semiempirical molecular orbital methods such as MNDO,¹ AM1,² PM3³ and successive methods,^{4,5} PDDG/PM3,⁶ and OMx,^{7,8} the self-consistent-charge density-functional tight-binding (SCC-DFTB) method is an alternative approximate approach derived from density functional theory (DFT) by neglect, approximation, and parametrization of interaction integrals.⁹ Although approximate methods are less accurate than DFT and *ab initio* methods on average, their main advantage is increased computational speed, which can be 2–3 orders of magnitude when compared to DFT and Hartree–Fock using medium-sized basis sets. This allows treating large molecules, a large number of conformers, and/or sufficiently long sampling for QM or QM/MM molecular dynamics simulations.¹⁰

The nonself-consistent version of DFTB^{11,12} and its basic integral approximations were proposed in the 1980s,^{13,14} still being the center around which all later extensions were developed. The DFTB energy, similar to other empirical tight-binding models, can be understood as a stationary approximation to the DFT functional¹⁵ in the spirit of the Harris functional approach.¹⁶ The methodology basically allows one to treat systems with small and large¹⁷ intramolecular charge transfer but fails for molecular systems with intermediate charge transfer. SCC-DFTB extends the DFTB method to charge self-consistency and can be derived by a second order expansion of the DFT total energy with respect to charge density fluctuations around a given reference charge density, usually chosen as a superposition of neutral atomic charge densities.⁹ The SCC-DFTB model now allows also for the treatment of systems with intermediate charge

transfer within a molecule and therefore has been a major step forward toward a generally applicable DFT-based semiempirical methodology. Several reviews have appeared concerning the basic formalism and selected applications;^{18–23} for a recent overview, we would like to point the readers also to a special issue of the *Journal of Physical Chemistry A*.²⁴

In recent years, several benchmark studies of SCC-DFTB appeared, showing the great success as well as the limitations of this method. Geometries are usually reproduced excellently.^{9,25–27} Similarly, relative energies of peptide conformers^{28–30} are nicely reproduced in comparison to higher level methods as well as hydrogen bonding energies.²⁶ While for reaction energies SCC-DFTB performs well on average,^{9,31} heats of formation are overestimated.^{25,26} Vibrational frequencies are reasonable, but severe failures have been noted for certain vibrational modes.^{26,31–35} A drawback inherited from the derivation of DFT is the missing dispersion interaction. An empirical correction has been suggested and shown to be crucial for the description of nucleic acid base stacking interactions³⁶ and the relative stability of α and 3_{10} helices in proteins.³⁷

The SCC-DFTB total energy consists of three terms:

$$E^{\text{SCC-DFTB}} = \sum_{iab} \sum_{\mu \in a} \sum_{\nu \in b} n_i c_{\mu i} c_{\nu i} H_{\mu\nu}^0 + \frac{1}{2} \sum_{ab} \Delta q_a \Delta q_b \gamma_{ab} + \frac{1}{2} \sum_{ab} V_{ab}^{\text{rep}} \quad (1)$$

the first term containing the DFTB matrix elements and the third one the DFTB repulsive potential. These two terms correspond to the non-self-consistent DFTB method,^{11,12} while the second term results from approximations of the second order term of the DFT

Received: November 26, 2010

Published: March 10, 2011

Taylor series expansion. Several limitations of the current formalism, which result from approximations inherent to those three terms, have been discussed recently,^{22,38} and current efforts to increase DFTB accuracy try to improve on these approximations. Recently, we have shown that a more sophisticated scheme for fitting the repulsive potential can also increase the overall accuracy to some degree.²⁷

In this work, we concentrate on extensions of the second order SCC term, leaving the other contributions, i.e., the first and the third terms, unchanged. As previous work has shown,^{22,38–40} extensions of the SCC contributions can improve the performance of SCC-DFTB for hydrogen bonded complexes and molecules with localized charges significantly, thereby improving the transferability of DFTB. These activities concern basically two recent developments, an improvement of the effective electron repulsion term in the SCC formalism, the γ function, and the extension to include third order terms.

The γ function describes the Coulomb interaction between atomic partial charges Δq_a . The functional form chosen for this interaction presupposes an inverse relation of atomic size and chemical hardness,⁹ which is true for elements within one row of the periodic table^{22,38} but not for elements of different periods. A particularly large deviation occurs for the hydrogen atom. A newly introduced γ^h function corrects this incorrect assumption using an extra term including one additional parameter and, as a result, systematically improves hydrogen bonding interactions.³⁹

The inclusion of approximate third order terms leads to a new degree of self-consistency.^{22,38} In SCC-DFTB, the Coulomb repulsion resulting from the charge density fluctuations as described by the second order SCC terms is computed in a monopole approximation utilizing a newly introduced parameter, the Hubbard parameter (chemical hardness). This parameter is computed from DFT for neutral atoms and is a constant for all charge states of the atom. While this approximation seems not to be problematic for many covalently bound systems, it is insufficient for molecules that contain large localized net charges. As has been shown, these systems require additional flexibility in the model; i.e., the Hubbard parameters have to become charge dependent, which is achieved by including the approximated third order terms.^{22,38}

The third order terms can be split up into two parts, a diagonal and an off-diagonal one. The diagonal terms lead to a charge dependent on-site self-interaction, the off-diagonal terms modify the SCC Coulomb repulsion between sites. The diagonal contributions significantly improve the proton affinities of CHNO-containing molecules, since in these calculations strongly localized net charges occur.³⁹ They also improve the proton affinities of phosphorus-containing molecules.⁴⁰ However, a reasonable accuracy was only achieved by adding an empirical energy contribution in a rather *ad hoc* fashion, which still did not lead to an acceptable transferability; i.e., different parameter sets had to be developed for different properties. Although these extensions have been shown to be important for describing proton affinities and hydrogen binding energies in various applications,^{23,41–47} further improvement is required to obtain a more transferrable method for general applications.

In the present study, we implement and test the off-diagonal third order contributions. In combination with the γ^h function and diagonal third order terms, this establishes a third generation of our DFTB methodology which will be called DFTB3. The off-diagonal terms are shown to overall improve the DFTB performance; most importantly, with this new formalism, a single set of parameters is able to reproduce many properties of CHNO- and

phosphorus-containing complexes with good accuracy. [A challenging problem still remains. The proton affinities for sp and sp hybridized nitrogen species are computed reasonably well; however, this is not the case for sp hybridized nitrogen systems, for which proton affinities are underestimated by about 10 kcal/mol. A pragmatic solution was suggested which introduced two nitrogen types, shifting the original N–H repulsive energy by these 10 kcal/mol for the second type.^{22,48} However, it remains unclear if this solution addresses the origin of the problem correctly. Another idea which we are currently exploring is to include d orbitals to nitrogen.]

In the next section, we give a short review of DFTB and SCC-DFTB as far as needed to explain the DFTB3 methodology. Next, computational details are discussed, including different ways for calculating proton affinities within the DFTB models. Finally, the performance of DFTB3 is evaluated for several test sets using data collected earlier^{39,40} and compared to SCC-DFTB and its previous γ^h and diagonal third order variants.

2. THEORETICAL APPROACH

The efficiency of DFTB is essentially linked to the use of a reference density ρ^0 , which is calculated from a superposition of neutral atomic densities ρ_a^0 . This allows one to compute Hamilton matrix elements in an atomic orbital (AO) basis in advance; i.e., no integral evaluation is necessary during the runtime of the calculation. The remaining contributions to the total energy are then approximated such that no further computational cost arises beyond the dominant step, which is the diagonalization of the precomputed Hamilton matrix. Therefore, all required approximations in DFTB are centered around the reference density ρ^0 and its deviation with respect to the DFT ground state density ρ , which is denoted by $\Delta\rho$. The approximations involved have been discussed in detail in previous publications.^{22,38,49} Essentially, the exchange-correlation energy contribution is expanded in a Taylor series expansion as

$$E^{\text{xc}}[\rho^0 + \Delta\rho] = E^{\text{xc}}[\rho^0] + \int \left[\frac{\delta E^{\text{xc}}[\rho]}{\delta \rho} \right]_{\rho^0} \Delta\rho + \frac{1}{2} \int' \int' \left[\frac{\delta^2 E^{\text{xc}}[\rho]}{\delta \rho' \delta \rho'} \right]_{\rho^0, \rho^0} \Delta\rho \Delta\rho' + \frac{1}{6} \int'' \int' \int' \left[\frac{\delta^3 E^{\text{xc}}[\rho]}{\delta \rho \delta \rho' \delta \rho''} \right]_{\rho^0, \rho^0, \rho^0} \Delta\rho \Delta\rho' \Delta\rho'' + \dots \quad (2)$$

where the abbreviations $\int = \int d^3r$, $\int' = \int d^3r'$, $\int'' = \int d^3r''$, $\rho = \rho(r)$, $\rho' = \rho(r')$, and $\rho'' = \rho(r'')$ are used. The total energy can then be written as

$$E[\rho^0 + \Delta\rho] = \sum_i n_i \left\langle \psi_i \left| -\frac{\nabla^2}{2} + V^{\text{ne}} + \int' \frac{\rho^{0'}}{|r-r'|} \right. \right. \\ + V^{\text{xc}}[\rho^0] \left. \right| \psi_i \rangle - \frac{1}{2} \int' \int' \frac{\rho^0 \rho^{0'}}{|r-r'|} - \int V^{\text{xc}}[\rho^0] \rho^0 + E^{\text{xc}}[\rho^0] \\ + E^{\text{nn}} + \frac{1}{2} \int' \int' \left(\frac{1}{|r-r'|} + \left. \frac{\delta^2 E^{\text{xc}}[\rho]}{\delta \rho \delta \rho'} \right|_{\rho^0, \rho^0} \right) \Delta\rho \Delta\rho' \\ + \frac{1}{6} \int'' \int' \int' \left. \frac{\delta^3 E^{\text{xc}}[\rho]}{\delta \rho \delta \rho' \delta \rho''} \right|_{\rho^0, \rho^0, \rho^0} \Delta\rho \Delta\rho' \Delta\rho'' + \dots \quad (3)$$

Here, n_i is the occupation number of the i th molecular orbital, V^{ne} is the nucleus–electron potential, V^{xc} is the exchange–correlation potential, and E^{nn} is the nucleus–nucleus repulsion.

Approximations of different levels of sophistication can be introduced by truncation of the Taylor series.³⁸ Standard (nonself-consistent) DFTB^{11,49} neglects second and higher order terms. This leads to a non-self-consistent scheme; i.e., the generalized eigenvalue problem has to be diagonalized only once. The SCC-DFTB method approximates the second order terms in the density fluctuations,⁹ while DFTB3 also includes the third order terms in an approximate way. In the following, a brief summary of SCC-DFTB and derivations of the third order terms as well as the γ^{h} function is given.

2.1. SCC-DFTB. The SCC-DFTB total energy is an approximation to the first three lines of eq 3. In a simple form the energy can be written as

$$E^{\text{SCC-DFTB}} = E^{\text{H0}} + E^{\gamma} + E^{\text{rep}} \quad (4)$$

First, $E^{\text{H0}} = \sum_{iab} \sum_{\mu \in a} \sum_{\nu \in b} n_i c_{\mu i} c_{\nu i} H_{\mu\nu}^0$ (see eq 1) is the energy contribution from an atomic orbital Hamiltonian depending only on the reference density. The determination of the atomic reference densities ρ_a^0 and the LCAO basis functions ϕ_{μ} needed for the calculation of $H_{\mu\nu}^0$ are discussed elsewhere.¹¹ The Hamilton and overlap matrix elements $H_{\mu\nu}^0$ and $S_{\mu\nu}$ are pre-computed and tabulated; i.e., they do not have to be computed during the runtime of the program. This and the use of a minimal valence basis set leads to huge computational savings (2–3 orders of magnitude) compared to full DFT.

Second, the repulsive energy contribution E^{rep} is an approximation of the so-called DFT double-counting terms, the core repulsion terms, and the exchange–correlation contributions of the second line of eq 3. In TB theory, these terms are usually approximated as a sum of one-center terms and short-ranged two-center potentials V_{ab}^{rep} :¹⁵

$$\begin{aligned} & -\frac{1}{2} \int' \int' \frac{\rho^0 \rho'^0}{|r-r'|} + E^{\text{xc}}[\rho^0] - \int V^{\text{xc}}[\rho^0] \rho^0 + E^{\text{nn}} \\ & \approx \sum_a V_a^{\text{rep}}[\rho_a^0] + \frac{1}{2} \sum_{ab} V_{ab}^{\text{rep}}[\rho_a^0, \rho_b^0, r_{ab}] \end{aligned} \quad (5)$$

where r_{ab} is the distance between atoms a and b . The atomic contributions are a constant energy shift which cancel when considering energy differences. For DFTB, the atomic contributions are neglected, and a repulsive energy E^{rep} is defined as

$$E^{\text{rep}} = \frac{1}{2} \sum_{ab} V_{ab}^{\text{rep}}[\rho_a^0, \rho_b^0, r_{ab}] \quad (6)$$

Third, the energy contribution E^{γ} is derived from the second order term of eq 3

$$\begin{aligned} E^{2\text{nd}} &= \frac{1}{2} \int' \int' \left(\frac{1}{|r-r'|} + \left. \frac{\delta^2 E^{\text{xc}}}{\delta \rho \delta \rho'} \right|_{\rho^0, \rho'^0} \right) \Delta \rho \Delta \rho' \approx E^{\gamma} \\ &= \frac{1}{2} \sum_{ab} \Delta q_a \Delta q_b \gamma_{ab} \end{aligned} \quad (7)$$

where $\Delta q_a = q_a - q_a^0$ is the net charge of atom a and γ is a function taking account of the electron–electron interaction. The γ function is given by the integral over a product of two normalized Slater-type spherical charge densities. We want to highlight two main properties of γ_{ab} which are described in detail

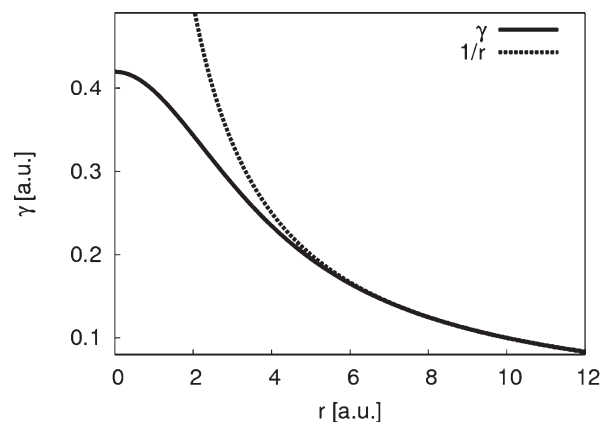


Figure 1. The γ function (solid line) plotted for the hydrogen–hydrogen interaction deviates from $1/r$ (dashed line) at short distances and yields the value of the Hubbard parameter $U_{\text{H}} = 0.4195$ au at $r = 0$ au.

in ref 9. For large distances r_{ab} , γ_{ab} basically reduces to $1/r_{ab}$; i.e., it describes a pure Coulomb interaction of the partial charges Δq_a and Δq_b . For $a = b$, γ_{ab} describes the on-site self-repulsion:

$$\gamma_{aa} = U_a \quad (8)$$

introducing the Hubbard parameter U_a (which is twice the chemical hardness). On the other hand, γ_{ab} imposes an inverse relationship between the Hubbard parameter and the covalent radius by⁹

$$\tau_a = \frac{16}{5} U_a \quad (9)$$

where τ_a is the exponent of the normalized Slater-type spherical charge density. Therefore, the Hubbard parameter affects two physical properties, the electron–electron interaction within one atom, i.e., the diagonal elements γ_{aa} , and the size of the atoms for estimating the two-center terms γ_{ab} . This estimated atomic size determines the deviation of γ_{ab} from $1/r_{ab}$, as shown in Figure 1. The Hubbard parameter U_a is the second derivative of the total energy of a single atom with respect to the occupation number of the highest occupied atomic orbital. In SCC-DFTB, it is estimated using Janack's theorem⁵⁰ by numerically calculating the first derivative of the energy of the highest occupied atomic orbital with respect to its occupation number for a neutral atom.

2.2. Third Order Term. An obvious extension of SCC-DFTB is to include also the third order term of the Taylor series expansion of the exchange correlation energy (eq 3). In second order DFTB, the chemical hardness of an atom (U_a) is constant irrespective of its charge state. For example, it does not allow anions to have a different chemical hardness value than the neutral atom or the cation. This is a severe limitation, as discussed in detail previously.^{22,38} Furthermore, in second order SCC-DFTB, the atoms are restricted to have a fixed shape as defined by the initial reference density ρ_a^0 . In third order, these restrictions are removed, which leads to a significant improvement for highly charged molecules.

The third order term as shown in eq 3 is given by

$$\begin{aligned} E^{3\text{rd}} &= \frac{1}{6} \int'' \int' \int' \left[\frac{\delta^3 E^{\text{xc}}[\rho]}{\delta \rho \delta \rho' \delta \rho''} \right]_{\rho^0, \rho'^0, \rho''^0} \Delta \rho \Delta \rho' \Delta \rho'' \\ &= \frac{1}{6} \int'' \int' \int' \frac{\delta}{\delta \rho''} \left[\frac{\delta^2 E^{\text{xc}}[\rho]}{\delta \rho \delta \rho'} \right]_{\rho^0, \rho'^0, \rho''^0} \Delta \rho \Delta \rho' \Delta \rho'' \end{aligned} \quad (10)$$

The same approximations as for the second order integrals can be applied^{38,39} (i.e., the description of the charge density fluctuations in terms of superposition of atomic contributions and the restriction of the charge density fluctuations to a monopole term, details see ref 9):

$$E^{3\text{rd}} \approx E^\Gamma = \frac{1}{6} \sum_{abc} \Delta q_a \Delta q_b \Delta q_c \left. \frac{d\gamma_{ab}}{dq_c} \right|_{q_a^0} \quad (11)$$

$$= \frac{1}{6} \sum_a \Delta q_a^3 \left. \frac{\partial \gamma_{aa}}{\partial q_a} \right|_{q_a^0}$$

$$+ \frac{1}{6} \sum_{a \neq b} \Delta q_a \Delta q_b \left(\left. \Delta q_a \frac{\partial \gamma_{ab}}{\partial q_a} \right|_{q_a^0} + \left. \Delta q_b \frac{\partial \gamma_{ab}}{\partial q_b} \right|_{q_b^0} \right) \quad (12)$$

Therefore, in the third order DFTB formalism, the derivative of the γ function with respect to charge introduces the desired chemical behavior for charged systems. For the diagonal terms (first term in eq 12), the derivative of γ implies via eq 8 a charge dependent Hubbard parameter (chemical hardness); i.e., the chemical hardness changes with charge state. Since U_a is also used to approximate the atom size in the damped Coulomb repulsion term γ , a charge dependent U_a will also make the atomic electron–electron repulsion charge dependent. For the off-diagonal terms (second term in eq 12), this effect applies for the electron–electron repulsion between two atoms. Note that γ_{ab} is dependent on the atomic charges only via the Hubbard parameters U_a and U_b . Introducing

$$\Gamma_{ab} = \left. \frac{\partial \gamma_{ab}}{\partial q_a} \right|_{q_a^0} = \left. \frac{\partial \gamma_{ab} \partial U_a}{\partial U_a \partial q_a} \right|_{q_a^0} \quad \text{with } a \neq b$$

$$\Gamma_{ba} = \left. \frac{\partial \gamma_{ab}}{\partial q_b} \right|_{q_b^0} = \left. \frac{\partial \gamma_{ab} \partial U_b}{\partial U_b \partial q_b} \right|_{q_b^0} \quad \text{with } a \neq b \quad (13)$$

$$\Gamma_{aa} = \left. \frac{\partial \gamma_{aa}}{\partial q_a} \right|_{q_a^0} = \frac{1}{2} \left. \frac{\partial \gamma_{aa} \partial U_a}{\partial U_a \partial q_a} \right|_{q_a^0}$$

where the latter definition is made to ease the summation; the third order energy contribution becomes

$$E^\Gamma = \frac{1}{6} \sum_{ab} \Delta q_a \Delta q_b (\Delta q_a \Gamma_{ab} + \Delta q_b \Gamma_{ba})$$

$$= \frac{1}{3} \sum_{ab} \Delta q_a^2 \Delta q_b \Gamma_{ab} \quad (14)$$

The derivative $(\partial \gamma_{ab})/(\partial U_a)$ can be computed analytically; details are given in the Supporting Information. The diagonal term $(\partial U_a)/(\partial q_a)|_{q_a^0}$ can be computed as the third derivative of the total energy of an atom with respect to charge. Practically, we compute the chemical hardness values for atoms in different charge states (applying Janack's theorem) and use these values to estimate the third derivative.³⁸

Thus, adding the approximated third-order contribution E^Γ to $E^{\text{SCC-DFTB}}$ (eq 1) yields the total energy of the third-order formalism:

$$E^{\text{DFTB3}} = E^{\text{H0}} + E^\gamma + E^\Gamma + E^{\text{rep}}$$

$$= \sum_{iab} \sum_{\mu \in a} \sum_{\nu \in b} n_i c_{\mu i} c_{\nu i} H_{\mu\nu}^0 + \frac{1}{2} \sum_{ab} \Delta q_a \Delta q_b \gamma_{ab}$$

$$+ \frac{1}{3} \sum_{ab} \Delta q_a^2 \Delta q_b \Gamma_{ab} + E^{\text{rep}} \quad (15)$$

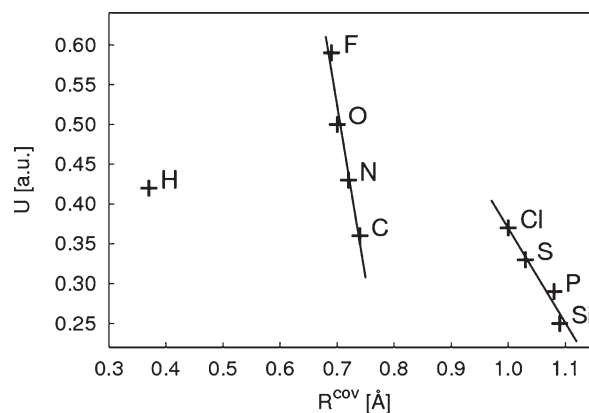


Figure 2. Calculated Hubbard parameters U versus covalent radii R^{cov} . The covalent radii are taken from the literature.⁵¹ For C, H, N, O, and F, values for R^{cov} are plotted that are estimated for bonds to second period elements; for Si, P, S, and Cl, values for R^{cov} are plotted that are estimated for bonds to third period elements. There is no overall inverse proportional relation as assumed by SCC-DFTB but only for elements within one period.

A detailed derivation of the Kohn–Sham equations, the third order Hamilton matrix elements

$$\sum_b \sum_{\nu \in b} c_{\nu i} (H_{\mu\nu} - \varepsilon_i S_{\mu\nu}) = 0 \quad \forall a, \mu \in a, i \quad (16)$$

$$H_{\mu\nu} = H_{\mu\nu}^0 + S_{\mu\nu} \sum_c \Delta q_c \left(\frac{1}{2} (\gamma_{ac} + \gamma_{bc}) + \frac{1}{3} (\Delta q_a \Gamma_{ac} + \Delta q_b \Gamma_{bc}) \right. \\ \left. + \frac{\Delta q_c}{6} (\Gamma_{ca} + \Gamma_{cb}) \right) \quad \forall a, b, \mu \in a, \nu \in b \quad (17)$$

and the force equations

$$F_{kx} = - \sum_{a \neq k} \sum_{\mu \in a} \sum_{\nu \in k} \sum_i n_i c_{\mu i} c_{\nu i} \left(2 \frac{\partial H_{\mu\nu}^0}{\partial R_{kx}} - 2 \varepsilon_i \frac{\partial S_{\mu\nu}}{\partial R_{kx}} \right. \\ \left. + \frac{\partial S_{\mu\nu}}{\partial R_{kx}} \left(\sum_c \Delta q_c \left(\gamma_{ac} + \gamma_{kc} + \frac{1}{3} (2 \Delta q_a \Gamma_{ac} + \Delta q_c \Gamma_{ca} \right. \right. \right. \\ \left. \left. \left. + 2 \Delta q_k \Gamma_{kc} + \Delta q_c \Gamma_{ck}) \right) \right) \right) - \Delta q_k \sum_{a \neq k} \Delta q_a \frac{\partial \gamma_{ak}}{\partial R_{kx}} \\ - \frac{1}{3} \Delta q_k \sum_{a \neq k} \Delta q_a \left(\Delta q_a \frac{\partial \Gamma_{ak}}{\partial R_{kx}} + \Delta q_k \frac{\partial \Gamma_{ka}}{\partial R_{kx}} \right) - \frac{\partial E^{\text{rep}}}{\partial R_{kx}} \quad \forall k, x \quad (18)$$

is provided in the Supporting Information.

2.3. The γ^{h} Function. The γ function represents the Coulomb repulsion between the density fluctuations within the DFTB approximation, i.e., for spherically constrained atomic densities. In ref 9, an analytical function has been derived, which is

$$\gamma_{ab} = \frac{1}{r_{ab}} - S(r_{ab}, U_a, U_b) \quad (19)$$

where S is a short-range function responsible for the correct convergence of γ_{ab} at $r_{ab} = 0$. This function imposes a simple rule, which implies that the chemical hardness of an atom is inversely proportional to its size.⁹ As has been pointed out

earlier, traditional semiempirical methods like MNDO, AM1, or PM3 use a similar approximation for the Coulomb interaction.³⁸ As discussed above, the Hubbard parameter U_a has a dual role: for the SCC on-site contributions, U_a models the effective Coulomb repulsion at site a , while for the off-diagonal terms, the inverse of U_a models the covalent radius of atom a ; i.e., it determines the deviation of γ_{ab} from $1/r_{ab}$. However, this inverse relation of chemical hardness and atomic size is not strictly valid across the periodic table;³⁸ it basically only holds within one period of the system of elements, as can be seen from Figure 2, which shows the

calculated Hubbard parameters for each element in dependence of the covalent radii. Therefore, in principle, a different γ_{ab} should be applied for different rows of the periodic table. Clearly, the deviation is the largest for hydrogen; therefore we proposed to modify γ_{ab} when hydrogen is involved and introduced a γ^h function as^{38,39}

$$\gamma_{ab}^h = \frac{1}{r_{ab}} - S(r_{ab}, U_a, U_b) \times h(r_{ab}, U_a, U_b) \quad (20)$$

where

$$h(r_{ab}, U_a, U_b) = \begin{cases} 1 & \text{if neither atom } a \text{ nor } b \text{ is of type hydrogen} \\ \exp\left[-\left(\frac{U_a + U_b}{2}\right)^\zeta r_{ab}^2\right] & \text{if at least one of atoms } a \text{ and } b \text{ is of type hydrogen} \end{cases} \quad (21)$$

In the following, we will refer to this function as the γ^h function in contrast to the γ function, as was used in standard SCC-DFTB ($h = 1$ for all cases). Note that, different than mentioned in ref 39, the γ^h function is also used for the H–H pair. The particular choice of h is to some degree arbitrary. On the other hand, its functional form is quite well physically motivated, correcting the shortcomings of the original function, since the chemical hardness of hydrogen simply cannot be used to represent the hydrogen covalent radius. We note that Clark and co-workers described a similar problem and modification for NDDO-based semiempirical methods.⁵²

Up to now, the parameters introduced in the second and third order extensions, in principle, can be calculated on the basis of DFT. Unfortunately, the parameter ζ in eq 21 cannot be computed from DFT but has to be fitted. However, as shown before,^{22,38} by choosing the parameter ζ such that the binding energy of the water dimer is reproduced correctly, γ^h becomes more repulsive in the covalent and hydrogen bonding region (see Figure 3) and improves hydrogen bonding systematically. We use the γ^h function in combination with the third order terms. Therefore, also the derivative of the γ^h function with respect to charge has to be calculated, which is shown in detail in the Supporting Information.

3. COMPUTATIONAL DETAILS

With DFTB3, one major difference from earlier studies^{39,40} is the way of calculating proton affinities, which we explain first. We continue with a description of different DFTB variants benchmarked in this work. A short review of the parameters of SCC-DFTB is given, and the new parameters of DFTB3 are introduced. Finally, we discuss problems that occur for nitrogen-containing species.

3.1. Calculation of Proton Affinities Using DFTB. The proton affinity is defined as the negative of the enthalpy change for the gas-phase reaction $A^-(g) + H^+(g) \rightarrow AH(g)$ at a given temperature. To avoid a large number of vibrational calculations, we consider in this work only the potential energy change and do not include zero-point correction, thermal contributions, and the PV term (difference between energy and enthalpy). This is done consistently for both reference calculations and DFTB

calculations. Due to the neglect of the one-center terms in the repulsive potential eq 5, the energy of a proton is not zero in DFTB⁵³ and can be computed in two ways:

- (i) First, it is given by the SCC-DFTB energy as (see eq 1):

$$E^{\text{SCC-DFTB}}(\text{H}^+) = \frac{1}{2}\gamma_{\text{HH}} = \frac{1}{2}U_{\text{H}} \quad (22)$$

This is a direct result of neglecting the one-center terms in the repulsive potential of eq 5, since eq 6 is used for all practical implementation and applications.⁵³ Therefore, the energy of the proton is given by half of the Hubbard parameter of hydrogen, which is 131.62 kcal/mol when computed using the DFT-PBE functional. This value may not be considered an accurate estimate since the Hubbard parameter is computed for the neutral hydrogen atom; however, it is consistent with the SCC-DFTB formalism.

- (ii) Alternatively, the one-center contribution to the repulsive potential can also be computed directly⁵³ as

$$V_{\text{H}}^{\text{rep}}[\rho_{\text{H}}^0] = E^{\text{DFT}} - E^{\text{SCC,el}} \quad (23)$$

With the energy of the hydrogen atom $E^{\text{DFT}} = -0.49772$ H (B3LYP/6-311++G(d,p)) and the electronic part of the SCC-DFTB energy, $E^{\text{SCC,el}} = E^{\text{H}^0} + E^{\gamma} = -0.27164$ H (first and second term in eq 1, here $E^{\gamma} = 0$ au) gives a one-center repulsive energy contribution for the hydrogen atom of $V_{\text{H}}^{\text{rep}}[\rho_{\text{H}}^0] = -141.87$ kcal/mol.⁵³ [Within the mio parameter, the spin-polarization energies are calculated with LDA; when using PBE values instead, the electronic energy contribution for the hydrogen atom is -0.27966 au, which gives $V_{\text{H}}^{\text{rep}}[\rho_{\text{H}}^0] = -136.83$ kcal/mol. For details, see ref 65.] For the proton, the energy within SCC-DFTB is then given by $0.5U_{\text{H}} + V_{\text{H}}^{\text{rep}}[\rho_{\text{H}}^0] = 10.25$ kcal/mol. Clearly, the electronic energy of a proton should be equal to zero; however, U_{H} is calculated as the derivative of the highest occupied atomic orbital with respect to the occupation number for the neutral hydrogen atom and cannot completely compensate for $V_{\text{H}}^{\text{rep}}[\rho_{\text{H}}^0]$ in the case of H^+ . For this unique situation where the total charge of the system is removed, the perturbative approach of SCC-DFTB fails. Therefore, the energy of

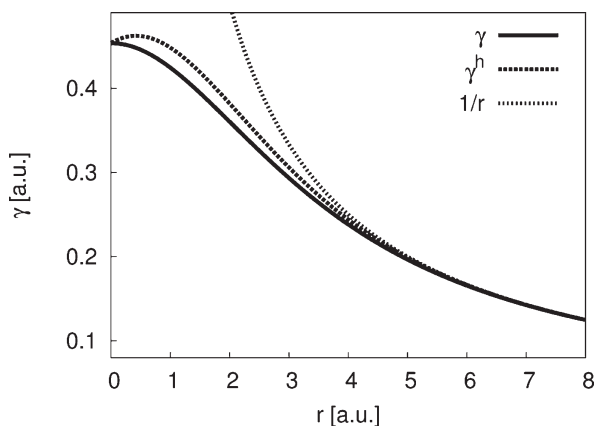


Figure 3. The γ^h function plotted for the OH pair ($U_O = 0.4954$ au, $U_H = 0.4195$ au). This function is more repulsive than the original γ function but still yields the same limits at $r = 0$ au and at $r \rightarrow \infty$.

the proton was set to $+141.87$ kcal/mol in earlier studies.^{39,53}

With DFTB3, the direct calculation (i) is slightly modified:

$$E^{\text{DFTB3}}(\text{H}^+) = \frac{1}{2} \Delta q_{\text{H}}^2 \gamma_{\text{HH}}^h + \frac{1}{3} \Delta q_{\text{H}}^3 \Gamma_{\text{HH}} = \frac{1}{2} U_{\text{H}} - \frac{1 \partial U_{\text{H}}}{6 \partial q_{\text{H}}} \Big|_{q_{\text{H}}^0} \quad (24)$$

The Hubbard derivative $(\partial U_{\text{H}})/(\partial q_{\text{H}})|_{q_{\text{H}}^0}$ is -0.1857 H using DFT-PBE, and the energy of the proton then equals 151.04 kcal/mol, now overestimating the value of $-V_{\text{H}}^{\text{rep}}[\rho_{\text{H}}^0]$.

In previous applications, approach ii has been used,^{39,40,53,54} however, this may not be the best choice for general applications. In principle, the energy of the proton is just a constant and not relevant when relative proton affinities are of interest, as, for example, for proton transfer reactions. However, it becomes important when absolute proton affinities and pK_{a} are of interest.⁵⁵ In the following, we will show that fitting the Hubbard derivatives leads to a drastic improvement of the performance of the method. When Hubbard derivatives are fitted, as in the current work, it is more consistent to use eq 24. Computing the PAs with

$$E^{\text{PA}} = E^{\text{A}^-} + E^{\text{H}^+} - E^{\text{AH}} \quad (25)$$

and using a fixed value for the DFTB energy of the proton E^{H^+} means that a change in the Hubbard derivative due to fitting affects only the energy of the molecule AH (E^{AH}) such that the reference proton affinity E^{PA} is obtained irrespective of the potential well depth of the A–H bond. This problem is resolved when calculating E^{H^+} with eq 24, E^{H^+} being different for different hydrogen Hubbard derivatives.

For this reason, we decide in the present work to consistently determine the energy of the proton with eqs 22 and 24 depending on the level of theory.

3.2. DFTB Variants. In DFTB, a Taylor series expansion is applied for the DFT exchange-correlation energy. While DFTB, the nonself-consistent variant, includes terms up to first order, SCC-DFTB includes also the second order term and DFTB3 also the third order term. For consistent naming, we decided to use the names DFTB, DFTB2, and DFTB3. Note that DFTB2 corresponds to the formally called SCC-DFTB. By default, the

Table 1. DFTB Variants Compared in Present Study

name	γ^h function	diagonal third order terms	off-diagonal third order terms
DFTB2	no	no	no
DFTB2- γ^h	yes	no	no
DFTB3- γ	no	yes	yes
DFTB3-diag	yes	yes	no
DFTB3	yes	yes	yes

standard γ function is used for DFTB2 and the γ^h function for DFTB3. In the following benchmark, we will compare

- DFTB2: formally called SCC-DFTB, using the standard γ function (eq 19) as derived in ref 9
- DFTB3: full third order extension (eq 15) including the γ^h function (eq 20) as derived in present work.

Starting from DFTB2, three major changes have been made to develop DFTB3. First, the γ^h function substitutes the standard γ function; second, diagonal third order terms are included; and third, off-diagonal third order terms are taken into account. To illustrate the effects of each of these extensions separately, we also present results for the following intermediate variants:

- DFTB2- γ^h : the standard γ function of DFTB2 is exchanged by the γ^h function
- DFTB3- γ : the standard γ function is used in connection with DFTB3
- DFTB3-diag: the γ^h function is used, and only the diagonal third order terms are included (second term in eq 12 is neglected) as suggested in refs 38 and 39.

Note that by introducing DFTB3 the intermediate variants become obsolete and are not recommended for practical applications. An overview of all DFTB variants is given in Table 1.

The additional computational costs of the γ^h function and the full third order extensions are negligible compared to the diagonalization of the Hamilton matrix, which is the time-limiting step in the DFTB methodology. Therefore, the computer time requirements are roughly the same for all variants in Table 1.

3.3. Parameter Sets. **3.3.1. Electronic and Repulsive Parameters.** The parameters for DFTB2 can be divided into two groups: atomic and diatomic parameters. A short summary of the different parameters is given in the following; for more details, see refs 9, 11, 38, 54, 56, and 57.

The atomic parameters are the two confinement radii for wave function and atomic reference density, the Hubbard parameter, and the spin-polarization energy. The last two parameters are computed from DFT and are not freely adjustable; the spin-polarization energy is only needed in order to compute heats of formation.²⁷ The two confinement radii are used for a proper choice of LCAO basis functions ϕ_{μ} and atomic reference densities ρ_{a}^0 . With these atomic parameters, one- and two-center integrals of the charge-independent part of the Hamiltonian H^0 are calculated in advance and tabulated; for more details, see ref 22.

The two-body potentials in E^{rep} (see eq 6) contain the diatomic parameters, which are usually fitted to reproduce reference data such as reaction energies and geometries of small molecules. For adequate fitting with several objectives, different techniques have been described in the literature.^{11,27,58,59} In this work, we use the mio parameters for all atoms and pairs including C, H, N, and O,^{9,57} which are available from www.dftb.org.

Table 2. Parameter Sets for the Different DFTB Variants

variant	set ^a	ζ^b	$U_C^{d,c}$	$U_H^{d,c}$	$U_N^{d,c}$	$U_O^{d,c}$	$U_P^{d,c}$	$U_S^{d,c}$
DFTB2								
DFTB2- γ^h	calc	3.70 ^d						
DFTB3- γ	calc		-0.1492	-0.1857	-0.1535	-0.1575	-0.0702	-0.0695
DFTB3-diag	calc	4.53	-0.1492	-0.1857	-0.1535	-0.1575	-0.0702	-0.0695
DFTB3-diag	fit	5.0	-0.04	-0.14	-0.11	-0.17	-0.07	-
DFTB3	calc	4.05	-0.1492	-0.1857	-0.1535	-0.1575	-0.0702	-0.0695
DFTB3	fit	4.2	-0.23	-0.16	-0.13	-0.19	-0.14	-

^a “calc” stands for calculated Hubbard derivatives U^d and/or ζ fitted to the water dimer and “fit” for a set of parameters fitted to a large set of binding energies and proton affinities. For details, see the text. ^b ζ is the unitless parameter as defined in eq 21. ^c $U_X^d = (\partial U_X)/(\partial n_X)|_{m_X}$ is the Hubbard derivative with respect to the occupation number of the highest occupied atomic orbital n_X of atom type X in atomic units. In our third order formalism, we define $(\partial U_X)/(\partial q_X)|_{q_X} = (\partial U_X)/(\partial n_X)|_{n_X}$. ^d Note that ζ is fitted to yield a binding energy for the water dimer of -4.9 kcal/mol in contrast to $\zeta = 4.5$ as reported in ref 39 for DFTB2- γ^h where ζ was fitted to minimize the error of 22 selected binding energies.

Additionally, we use phosphorus parameters as described in ref 40, which have not been released up to now, since the performance was not satisfactory at the second order DFTB2 level of theory. Many of these problems are resolved with DFTB3; however, since the repulsive potentials have been determined for DFTB2, these parameters are still not satisfactory for all purposes, as shown in detail below. Clearly, a new parametrization for DFTB3 has to be developed; nevertheless, current parameters provide reasonable geometries for a wide range of molecules, and they will be available soon at www.dftb.org.

Briefly, the confinement radii for phosphorus are chosen as those for sulfur,⁶⁰ that is, $3.8 a_0$ for the wave function of the 3s and 3p valence orbitals, $4.4 a_0$ for 3d orbitals, and $9.0 a_0$ for the density compression. While the DFT eigenvalue of the d orbital is calculated as $\varepsilon_d = 0.02$ H, it was set to $\varepsilon_d = 0.52$ H in order to reduce excessive d-orbital involvement in binding situations. The repulsive potentials for six different pairs (P–P, P–C/H/N/O/S) are fitted to a B3LYP/6-31G(d) reference and are truncated to zero in the range of 1.7 – 3.3 Å using the molecules PH₃, PCH, HPCH₂, H₂PCH₃, PN, HPNH, H₂PNH₂, P₂, HPPH, H₂PPH₂, OPH, H₃PO₄, H₄PO₅⁻, HPS, and H₂PSH. Details for the general fitting procedure, as has been carried out for the phosphorus parameters, can be found in refs 9 and 57.

3.3.2. New Parameters. The γ^h function for the pairs HX ($X \in \{C, H, N, O, P, S\}$) describes the dependence between the size of the atom and the electron–electron interaction more correctly; one additional, purely empirical parameter ζ is necessary. It can be determined using only one data point, the binding energy of the water dimer, for which the most accurate theoretical value is 5.0 kcal/mol using CCSD(T).⁶¹ Nevertheless, to stay consistent within our fitting procedure as described below, we choose the similar value of 4.9 kcal/mol, which is the result from the G3B3^{62,63} method. In Table 2, we denote this way of determining ζ as “calc”. We will also fit this parameter to reproduce an extended data set, in combination with fitting the Hubbard derivatives, then denoted as “fit”.

The third order Taylor series expansion of the exchange correlation energy makes use of the Hubbard derivatives $U_X^d = (\partial U_X)/(\partial q_X)|_{q_{OX}}$ which means one additional parameter per element. These can be determined by taking the numerical derivative of the corresponding Hubbard parameter of a neutral atom with respect to the occupation number of the highest occupied atomic orbital. In Table 2, the Hubbard derivatives are summarized and abbreviated as “calc”; they are calculated with the PBE exchange–correlation functional⁶⁴ and our in-house program TWOCENT.

Therefore, one parameter set we provide is the “calc” set, where only one parameter (ζ) is fitted to one system (water dimer), and all Hubbard derivatives are calculated. In a different approach, we fit all parameters for a large set of molecules, resulting in the parameter set “fit”. This has been done first for the DFTB3-diag method in ref 39.

It is important to note that fitting of the Hubbard derivatives and ζ basically affects hydrogen bonds and proton affinities; most properties of neutral molecules like equilibrium geometries are not significantly altered. Nevertheless, one has to be careful not to correct at the third order level for errors that result from the second order formalism, i.e., shortcomings resulting from the electronic and repulsive parameters of the original DFTB2. Our results indicate that the approximations in the third order terms account very well for the physical effects arising from that level. The remaining errors in the description of H-bonding and proton affinities seem not to result from the third order approximations but from the underlying second order DFTB2. For the moment, we optimize DFTB3 parameters to make it applicable to important chemical and biological problems without refining the DFTB2 approximations.

The idea of DFTB2 is to use as many parameters calculated from DFT as possible. By fitting the Hubbard derivatives U_X^d we are leaving this spirit, and it seems natural to challenge the insistence on DFT-calculated Hubbard parameters U_X . Surely, a fitting also of these parameters may lead to an improved chemical accuracy, however, at cost of the following benefits: first, a physically robust and transferable method; second, an easy detection of systematic errors; and third, a small space of parameters allowing an easier fitting of the remaining parameters (especially because the Hubbard parameters affect mainly all chemical properties for systems of biological relevance). Please also note that the third order formalism introduces new physics into our method that cannot be compensated for just by a new set of fitted Hubbard parameters. In that sense, our future work is focused on avoiding empirical fitting of Hubbard parameters and derivatives by improving the electronic (confining radii) and repulsive parameters, that have been until now fitted on DFTB2, at the DFTB3 level.

To optimize the parameters for DFTB3-diag, Yang et al. chose a weighted penalty function where the properties of interest included binding energies, proton affinities, and the root-mean-square gradient of the included molecules calculated at the reference structure.³⁹ Finally they minimized the penalty function using a genetic algorithm optimizing the Hubbard derivatives and the ζ parameter. In this work, we use a “brute force”

Table 3. Mean Unsigned and Maximum Absolute Deviation of Geometrical Properties^a of the G2 Set for 61 CHNO-Containing Closed Shell Molecules for Different DFTB Variants^b

parameter set ^b	N ^c	DFTB2		DFTB3				PBE ^d	B3LYP ^d	
		γ	γ^h	γ	diag		full			
				calc	fit	calc	calc	fit		
r (Å)	223	0.014	0.014	0.014	0.014	0.014	0.014	0.014	0.009	0.004
r^{\max} (Å)	223	0.065	0.067	0.061	0.064	0.064	0.062	0.063	0.060	0.041
a (deg)	187	0.9	0.9	0.9	0.9	0.9	0.9	1.0	0.4	0.4
a^{\max} (deg)	187	4.7	6.4	4.9	6.4	6.2	6.6	6.5	1.9	1.9

^aBond lengths, r ; bond angles, a ; max stands for maximum absolute deviation. Geometric data is compared to the MP2/cc-pVTZ calculations. For details, see the Supporting Information. ^bFor explanations, see Tables 1 and 2. ^cNumber of comparisons. ^dBasis set 6-311G(2d,2p).

fitting. A small set of parameters around the calculated values (ζ fitted to the water dimer) is chosen, and the performance is evaluated by calculating the mean unsigned error (MUE) of proton affinities and binding energies using geometry optimized molecules for each parameter set. Whenever the parameter set that performs the best reaches a boundary of the current range of parameters, the range is extended. The latter step is repeated until the best set does not reach any boundaries of the current range.

For the fitting of ζ and the Hubbard derivatives of C, H, N, and O, a set of 22 binding energies and 32 proton affinities as compiled by Yang et al.³⁹ is used to represent important biological properties. The calculations are carried out in the gas phase at 0 K without including the zero-point corrections for both reference and DFTB. Subsequently, the Hubbard derivative of phosphorus is fitted to a set of 18 proton affinities of phosphorus-containing molecules (compilation from ref 40) in the same manner but keeping all other parameters fixed. All molecules involved in the fitting procedure are listed in the following subsections.

It is found that the Hubbard derivative of carbon becomes very small during the fitting of DFTB3/fit, while all other Hubbard parameters stay close to the calculated values. To avoid getting unphysical values, we limit U_C^d to a lower boundary of -0.23 H. [For $U_C^d < -0.40$ H, we find that the self-consistent procedure does not converge for several molecules in our training set.] Similarly, U_C^d becomes quite large during the fit of DFTB3-diag/fit such that we limit it to an upper boundary of -0.04 au. Note that the fitted parameters are different from the ones published by Yang et al.³⁹ since the way of computing proton affinities is different (for details, see above).

The additional off-diagonal terms within DFTB3 seem to be more repulsive in comparison with DFTB3-diag; therefore, ζ becomes smaller to compensate for that, as shown in Table 2.

3.3.3. Nitrogen Hybridization: A Problematic Case for a Minimal Basis Set Method. Nitrogen hybridization seems to pose a problem for minimal basis set methods like DFTB as well as for NDDO type semiempirical methods.⁵² This problem, which may be related to the neglect of d orbitals in the basis set, is not corrected for by either the γ^h function or the third order terms and leads to dramatic errors when computing deprotonation energies. In previous studies,^{48,22} consistent errors of about 10 kcal/mol were found specifically for proton affinities of sp^3 hybridized nitrogen atoms. Therefore, a modified parameter set “NHmod” was introduced in which the N–H repulsive potential was shifted to correct for these errors. However, since sp^2 hybridized nitrogen atoms seem to be described correctly, this

correction has only to be applied for a certain electronic configuration of N. Therefore, similar to the situation in force fields, different “atom types” for N have to be introduced at the moment, which clearly limits DFTB’s applicability since these atom types are not allowed to change during a reaction. In this work, we present results for a “NHorg” and a “NHmix” parameter set. NHorg denotes the parameters for N–H bonds from the mio set; i.e., in this set, no different atom types occur. For the NHmix set, the mio potential is only used for compounds containing sp^2 or sp^1 nitrogen, whereas NHmod is applied for sp^3 hybridized nitrogen atoms. For reactions where a nitrogen changes its hybridization state from sp^2 or sp to sp^3 , the NHorg repulsive potential is used in order to have consistent energetic contributions for the N–H atom pairs. [Note, different than in present work, Yang et al.³⁹ defined NHmix in a way that NHmod is also used for calculating the proton affinity of NH_2^- . Since the orbitals calculated on the NH_2^- molecule look similar to orbitals on sp^2 nitrogen, we apply NHorg for that case.]

The fitting procedure for ζ and the Hubbard derivatives is applied separately for NHorg and NHmix; however, both optimized parameters turn out to be equal. This extends the transferability of the “fit” parameter sets (see Table 2) and implies that, in addition to the case for “calc”, in the case for “fit” the NHorg and NHmix results differ only for test molecules where a sp^3 nitrogen is bound to hydrogen.

4. BENCHMARKS AND DISCUSSION

In the following subsections, we present benchmark calculations for the different DFTB variants shown in Table 1 regarding geometries, binding energies, proton affinities, and proton transfer barriers for CHNO-containing molecules and also compare the results with commonly used density functionals. We further show results on proton affinities and hydrolysis reactions of phosphorus-containing molecules. Finally, some general benchmarks are provided for phosphorus parameters.

The parameters used for the γ^h function and third order terms are given in Table 2, and if not explicitly stated, the NHorg repulsive potential is used. Binding energies, proton affinities, proton transfer barriers, and reaction energies are computed using the potential energies at 0 K without including any zero-point energy correction. Deviations are given as the difference of high level ab initio methods ($E^{\text{method}} - E^{\text{high-level}}$), where the high level calculations are performed using the Gaussian 03 program.⁶⁵

The compilation and notation for binding energies and proton affinities are taken from ref 39, proton affinities and hydrolysis reactions of phosphorus-containing molecules from ref 40.

Table 4. Deviation of DFTB in Comparison to B3LYP/cc-pVTZ for Selected Bond Lengths r in Å

parameter set ^a	B3LYP	DFTB2		DFTB3				
		γ	γ^h	γ	diag		full	
				calc	calc	fit	calc	fit
rCC in CH ₃ COO ⁻	1.567	+0.047	+0.051	+0.000	+0.011	-0.003	+0.005	-0.004
rHO in OH ⁻	0.971	+0.033	+0.011	+0.005	-0.010	-0.009	-0.003	-0.003
rOH in (H ₂ O) ₂ ^b	1.945	-0.056	-0.122	-0.060	-0.125	-0.120	-0.116	-0.117

^aFor explanations, see Tables 1 and 2. ^brOH: hydrogen bond length in water dimer.

Table 5. 22 Binding Energies in kcal/mol: Deviation of DFTB in Comparison to G3B3^a

parameter set ^b	G3B3	DFTB2		DFTB3				
		γ	γ^h	γ	diag		full	
				calc	calc	fit	calc	fit
2H ₂ O	-4.9	+1.6	-0.0	+1.5	-0.0	+0.2	-0.0	+0.0
3H ₂ O	-15.1	+5.5	-0.6	+5.4	-0.5	+0.2	-0.3	-0.1
4H ₂ O	-27.4	+9.7	+0.6	+9.4	+0.8	+1.8	+0.8	+1.1
5H ₂ O	-36.3	+13.3	+1.4	+12.5	+1.8	+3.0	+1.3	+1.7
2H ₂ O(H ⁺)	-33.9	+4.5	-2.0	+5.9	+2.4	+3.4	+0.9	+2.1
3H ₂ O(H ⁺)	-57.3	+10.4	-0.1	+11.6	+5.3	+6.6	+3.7	+5.3
4H ₂ O(H ⁺)	-77.2	+13.9	+1.1	+15.0	+6.4	+7.9	+5.0	+6.7
5H ₂ O(H ⁺)	-91.9	+18.3	+1.8	+19.7	+7.2	+9.1	+6.2	+8.1
2H ₂ O(-H ⁺)	-27.4	-5.1	-12.8	+1.5	-3.4	-1.5	-5.9	-3.2
3H ₂ O(-H ⁺)	-48.6	-2.6	-17.0	+5.3	-6.5	-3.8	-8.4	-5.3
4H ₂ O(-H ⁺)	-66.7	+0.3	-17.5	+9.0	-5.0	-1.8	-7.2	-3.5
5H ₂ O(-H ⁺)	-86.3	+6.1	-18.2	+14.2	-7.7	-4.1	-7.8	-4.7
NH ₃ (H ₂ O) ^c	-6.6	+3.2	+2.1	+3.1	+2.0	+2.2	+2.1	+2.1
NH ₄ ⁺ (H ₂ O) ^c	-20.4	+0.6	-3.4	+1.4	-1.1	-0.7	-1.3	-0.9
6H ₂ O_book	-45.8	+16.7	+1.2	+16.5	+1.5	+3.2	+1.7	+2.2
6H ₂ O_cage	-46.6	+17.2	+0.3	+17.6	+0.3	+2.1	+1.5	+1.8
6H ₂ O_prism	-47.2	+17.6	-0.0	+18.0	+0.1	+2.0	+1.3	+1.7
6H ₂ O_ring	-44.7	+16.5	+1.8	+15.3	+2.4	+3.9	+1.5	+2.1
methylimidazole(-H ⁺)(H ₂ O)	-15.9	+4.1	+2.0	+3.2	+1.4	+1.5	+1.2	+1.1
methylimidazole(H ₂ O)_1	-6.2	+2.4	+1.4	+2.6	+1.8	+1.9	+1.9	+2.0
methylimidazole(H ₂ O)_2	-8.2	+3.5	+2.6	+2.8	+2.0	+2.2	+1.9	+1.9
methylimidazoleH ⁺ (H ₂ O)	-16.0	+3.3	+1.2	+3.9	+2.2	+2.3	+2.3	+2.5
MUE		8.0	4.0	8.9	2.8	3.0	2.9	2.7
MSE		+7.3	-2.5	+8.9	+0.6	+1.9	+0.1	+1.1
MAX		18.3	18.2	19.7	7.7	9.1	8.4	8.1

^aThe binding energy is computed as the difference between the complex and the isolated molecules at 0 K; no zero-point energy correction has been included. For the DFTB methods, the deviation is given as the difference of the G3B3 method ($E^{\text{method}} - E^{\text{G3B3}}$). Compilation of molecules and notation taken from ref 39. Examples of notation: “2H₂O”, neutral water dimer; “2H₂O(H⁺)”, protonated water dimer; “2H₂O(-H⁺)”, deprotonated water dimer; “6H₂O_book”, neutral water hexamer in the book configuration; “methylimidazole(-H⁺)(H₂O)”, deprotonated methylimidazole complexed with water; “methylimidazole(H₂O)_1”, neutral methylimidazole complexed with water as hydrogen-bond donor; “methylimidazole(H₂O)_2”, neutral methylimidazole complexed with water as the hydrogen-bond acceptor; “methylimidazoleH⁺(H₂O)”, protonated methylimidazole complexed with water. ^bFor explanations, see Tables 1 and 2. ^cWhen applying NHmix, the results are slightly but not significantly different.

4.1. Geometries. The performance of the different DFTB variants is tested for the charge-neutral closed-shell molecules of the G2⁶⁶ set. As shown in Table 3, the geometries do not change significantly for all tested DFTB variants and parameter sets. Similarly, the different NH repulsive potentials NHorg and NHmix cause only very small differences for geometries; for details, see the Supporting Information.

Significant differences occur for charged molecules; some of them are summarized in Table 4. For example, the C–C bond length in the acetate anion is overestimated by DFTB2 in comparison to B3LYP^{67–69}/cc-pVTZ,⁷⁰ that error becomes smaller for the DFTB variants including third order terms. Similar findings are obtained for the O–H bond length of the hydroxide anion, even though in this case also the γ^h function has

Table 6. 23 Proton Affinities with Acidic Oxygen in kcal/mol: Deviation of DFTB in Comparison to G3B3^a

parameter set ^b	G3B3	DFTB2		DFTB3				
		γ	γ^h	γ	diag		full	
				calc	calc	fit	calc	fit
H ₂ O	398.4	+16.3	+18.5	+8.0	+5.8	-1.6	+7.5	-1.8
2H ₂ O	375.9	+9.6	+5.7	+8.0	+2.3	-3.2	+1.7	-5.1
3H ₂ O	365.0	+8.1	+1.9	+7.7	-0.4	-5.6	-0.7	-7.1
4H ₂ O	359.1	+7.0	+0.3	+7.5	-0.1	-5.2	-0.5	-6.5
5H ₂ O	348.4	+9.2	-1.1	+9.6	-3.7	-8.7	-1.6	-8.3
CH ₃ OH	392.6	-5.7	-2.6	+3.3	+5.8	-0.7	+5.9	-0.3
CH ₃ CH ₂ OH	388.3	-1.5	+1.6	+6.5	+9.3	+2.6	+9.0	+2.2
CH ₃ CH ₂ CH ₂ OH	387.6	-2.2	+1.0	+6.0	+8.7	+1.9	+8.6	+2.0
CH ₃ -CH(OH)-CH ₃	385.6	+1.4	+4.7	+8.2	+11.7	+4.6	+10.7	+3.3
HCOOH	351.2	+1.7	+3.4	+8.6	+14.2	+7.1	+10.0	+2.9
CH ₃ COOH	355.1	+1.1	+3.2	+6.8	+12.7	+5.6	+8.5	+0.6
CH ₃ CH ₂ COOH	354.5	+1.0	+3.4	+7.5	+13.1	+6.0	+9.3	+1.5
C ₆ H ₅ OH	356.7	-4.7	-2.4	+8.0	+11.0	+5.2	+9.7	+4.0
<i>p</i> -CH ₃ -C ₆ H ₄ OH	357.9	-5.6	-3.1	+7.4	+10.5	+4.5	+9.2	+3.7
<i>p</i> -NO ₂ -C ₆ H ₄ OH	334.6	-9.3	-7.5	+2.2	+5.2	-0.7	+3.5	-1.3
H ₃ O ⁺	171.2	-0.4	-4.7	+10.6	+9.0	+5.6	+6.3	+4.3
2H ₂ O(H ⁺)	200.2	-3.3	-2.7	+6.3	+6.6	+2.3	+5.4	+2.2
3H ₂ O(H ⁺)	213.4	-5.3	-5.1	+4.4	+3.2	-0.9	+2.3	-1.2
4H ₂ O(H ⁺)	221.1	-4.7	-5.3	+4.9	+3.3	-0.7	+2.0	-1.4
5H ₂ O(H ⁺)	226.7	-5.3	-5.1	+3.5	+3.8	-0.4	+1.4	-1.9
CH ₃ OH ₂ ⁺	186.8	-8.3	-10.3	+6.5	+6.1	+2.0	+4.6	+2.2
H ₂ COH ⁺	177.1	-11.8	-13.8	+4.3	+4.3	+0.5	+2.6	-0.2
CH ₃ CHOH ⁺	190.2	-10.1	-10.8	+5.8	+6.5	+2.4	+5.1	+2.0
MUE		5.8	5.1	6.6	6.8	3.4	5.5	2.9
MSE		-1.0	-1.3	+6.6	+6.5	+1.0	+5.2	-0.2
MAX		16.3	18.5	16.6	14.2	8.7	10.7	8.3

^a The molecules are given in the protonated form. The proton affinity is computed with the potential energies at 0 K without any zero-point energy correction. For the DFTB methods, the deviation is given as the difference of the G3B3 method ($E^{\text{method}} - E^{\text{G3B3}}$). The compilation of the molecules is taken from ref 39. ^b For explanations, see Tables 1 and 2.

a significant effect. The hydrogen bond length in the water dimer is overestimated using B3LYP/cc-pVTZ⁷¹ due to the admixture of HF exchange; it is shorter for a pure GGA functional like PBE/cc-pVTZ, where this bond length is 1.917 Å. DFTB2 underestimates this bond length (1.889 Å), indicating that the Pauli repulsion may be underestimated by DFTB. Inclusion of the γ^h function even further shortens the hydrogen bond. It is important to note that this is a general trend (also valid for e.g. water clusters); i.e., hydrogen bond lengths are predicted systematically too short by DFTB.

4.2. Binding Energies. In a previous study, it has been shown that DFTB2 underestimates the strength of hydrogen bonding interactions.³⁹ The performance for hydrogen bonds is drastically improved using the γ^h function, as shown in Table 5, while the third order corrections alone (third) do not seem to have a substantial effect on these properties. However, the errors for the negative charged species are now more consistent with the ones of neutral and positive charged systems. The combination of both extensions in DFTB3-dia and DFTB3 adopts both improvements; the mean unsigned error in comparison to G3B3^{62,63} drops from 8 kcal/mol for DFTB2 to about 3 kcal/mol irrespective of the set of Hubbard derivative parameters (U^d) used. In ref 39, the test of DFTB3-

diag has been extended to a larger test set, and we expect similar results for DFTB3.

In many biological applications, DFT methods with medium-sized basis sets are applied. In order to compare DFTB with DFT, we compile also binding energies for the same molecule set (Table 5) using PBE and B3LYP with the 6-31+G(d,p) basis set, which give a mean unsigned error of 7.0 and 3.7 kcal/mol (for details, see the Supporting Information). These errors are significantly larger when using basis sets without a diffuse function. This of course is due to the basis set superposition error (BSSE), which can be remediated when including the counterpoise correction,^{72,73} dropping the MUE to 3.7 and 1.3 kcal/mol, respectively. Nevertheless, we think it is important to be aware of these large errors, for example, when studying larger biomolecular systems where the counterpoise correction is rarely done. Therefore, although it is often claimed that certain DFT functionals perform well for hydrogen bonding,⁷⁴⁻⁷⁷ this is only true for converged basis sets, which are often not used in practical applications. In such cases, the use of a well calibrated approximate method like DFTB can be an even more appropriate choice. For example, the finding that the active site of bacteriorhodopsin is scrambled using QM/MM-CPMD simulation may be related to an imbalanced description of QM, QM/MM, and MM interactions, where one factor

Table 7. Nine Proton Affinities with Acidic Nitrogen in kcal/mol: Deviation of DFTB and the NHorg Parameter Set in Comparison to G3B3^a

parameter set ^b	G3B3	DFTB2		DFTB3				
		γ	γ^h	γ	diag		full	
				calc	calc	fit	calc	fit
HCNH ⁺	176.0	-12.4	-14.6	+4.5	+4.3	+0.4	+2.8	+0.2
CH ₃ CNH ⁺	192.3	-14.3	-15.4	+2.9	+2.6	-1.2	+1.9	-0.9
C ₅ H ₅ NH ⁺	229.5	-17.1	-18.3	+1.3	+0.9	-3.5	+0.4	-2.1
methylimidazoleH ⁺	237.3	-12.7	-13.4	+5.3	+5.1	+0.8	+4.7	+2.1
methylguanidineH ⁺	249.3	-12.0	-13.4	+0.8	+0.4	-2.2	-0.8	-2.9
NH ₃	413.9	+10.4	+10.9	-0.9	-16.8	-0.0	-5.3	-0.2
NH ₄ ⁺	212.3	-24.4	-30.5	-9.2	-13.0	-15.0	-14.4	-15.8
CH ₃ NH ₃ ⁺	223.3	-26.8	-30.5	-10.2	-11.7	-15.1	-13.3	-15.3
1-aminobutaneH ⁺	228.2	-26.7	-29.9	-9.8	-11.4	-14.9	-12.6	-14.6
MUE		17.4	19.7	5.0	7.4	5.9	6.2	6.0
MSE		-15.1	-17.2	-1.7	-4.4	-5.6	-4.1	-5.5
MAX		26.8	30.5	10.2	16.8	15.1	14.4	15.8

^aThe molecules are given in the protonated form. The proton affinity is computed with the potential energies at 0 K without any zero-point energy correction. For the DFTB methods, the deviation is given as the difference of the G3B3 method ($E^{\text{method}} - E^{\text{G3B3}}$). The compilation of the molecules is taken from ref 39. ^bFor explanations, see Tables 1 and 2.

contributing to the imbalance may be BSSE.⁷⁸ The application of empirical dispersion corrections would even worsen the problem, since dispersion further strengthens the interaction, i.e., leads to an even larger overbinding.

4.3. Proton Affinities. As shown in earlier studies,^{39,79} DFTB2 overestimates proton affinities (PA) that implicate acidic oxygen. Yang et al. report an improvement with DFTB3-dia for molecules in which the charge is strongly localized, a situation where the third-order term contributes accordingly. In these studies, the DFTB2 energy of the proton was assumed to be 141.9 kcal/mol; in the present work, we use eq 22 for DFTB2 and eq 24 for DFTB3. Consequently, the proton affinities as compiled in Table 6 are shifted by about 10 kcal/mol for DFTB2 in comparison to the earlier studies.

While the mean signed error (MSE) for DFTB2 in comparison to G3B3 is quite small, the proton affinities of negatively charged molecules are overestimated and the proton affinities for neutral molecules underestimated. This holds true also when including the γ^h function. The situation changes when looking at the third order variants. Even though the MUE is not significantly reduced (or even enlarged) in comparison to DFTB2, the proton affinities for almost all molecules are consistently overestimated, and the MSE is (almost) as large as the MUE (+5.2 vs 5.5 kcal/mol in the case of DFTB3/calc). This indicates a consistent overbinding of the O–H bond. This error, however, is not related to the third order formalism but has its roots already in the repulsive potential of DFTB2. As Otte et al.²⁶ mentioned, the O–H bond shows an overbinding of about 6–7 kcal/mol. This overbinding can also be roughly estimated by half of the atomization energy error of H₂O, which is 5.8 kcal/mol for DFTB3/calc (using PBE spin-polarization energies; for details, see ref 27) in comparison to G3B3. This value is very similar to the MSE of DFTB3/calc (5.2 kcal/mol) in Table 6, leading to the conclusion that removing this overbinding remedies the error for the proton affinities. Indeed, once fitting the third order and γ^h function parameters (DFTB3/fit), the MSE can be removed to obtain a MUE as small as 2.9 kcal/mol. [For water clusters, we

note that the PA can be written as the sum of PA for a (neutral or protonated) water molecule and the difference in the binding energies of water clusters of different protonation states. Therefore, the errors in the water cluster PAs can be understood in terms of the errors in the PA of a single (neutral or protonated) water and errors in the binding energies of the relevant water clusters. For example, the fairly large error for the PA of a neutral (H₂O)₅ is due mainly to the fact that DFTB3/fit overestimates the binding energy of a deprotonated (H₂O)₅ (-4.7 kcal/mol, see Table 5) but slightly underestimates the binding energy of a neutral (H₂O)₅ (+1.7 kcal/mol, see 4).] This would not work for DFTB2, indicating that the third order terms systematically lead to an improvement of DFTB.

Proton affinities with acidic nitrogen are shown in Table 7. Here, DFTB2 shows large errors, which are systematically improved by all third order variants. Large errors remain for the last three molecules in Table 7 with sp³ nitrogen, which show a systematic error of more than 10 kcal/mol, as discussed in detail already in ref 39. The use of NHmod specifically for sp³ hybridized nitrogen, although not satisfactory from a theoretical point of view, remedies this problem (see Table 8). That way, the remaining MUE for DFTB3/calc/NHmix is only 2.5 kcal/mol.

Another encouraging result is the improvement of the proton affinity for NH₂⁻. While for DFTB2 the proton affinity is overestimated, it is underestimated for DFTB3-dia. The error is then substantially reduced using the full third order variants, DFTB3- γ and DFTB3, showing the first example where the third order off-diagonal terms seem to be of importance.

Due to the hybridization problem, the error analysis for the N–H bond is more involved. Nevertheless, the overbinding of the N–H bond calculated as a third of the error in the atomization energy of NH₃ for DFTB3/calc as compared to G3B3 is 2.9 kcal/mol (using PBE spin-polarization energies, details see ref 27), which is comparably small. With the O–H overbinding of 5.8 kcal/mol, we can estimate the error for the relative proton affinity between oxygen- and nitrogen-containing molecules to be roughly (5.8–2.9) kcal/mol = 2.9 kcal/mol,

Table 8. Nine Proton Affinities with Acidic Nitrogen in kcal/mol: Deviation of DFTB and the NHmix Parameter Set in Comparison to G3B3^a

parameter set ^b	G3B3	DFTB2		DFTB3				
		γ	γ^h	γ	diag		full	
				calc	calc	fit	calc	fit
HCNH ⁺	176.0	-12.4	-14.6	+4.5	+4.3	+0.4	+2.8	+0.2
CH ₃ CNH ⁺	192.3	-14.3	-15.4	+2.9	+2.6	-1.2	+1.9	-0.9
C ₅ H ₅ NH ⁺	229.5	-17.1	-18.3	+1.3	+0.9	-3.5	+0.4	-2.1
methylimidazoleH ⁺	237.3	-12.7	-13.4	+5.3	+5.1	+0.8	+4.7	+2.1
methylguanidineH ⁺	249.3	-12.0	-13.4	+0.8	+0.4	-2.2	-0.8	-2.9
NH ₃	413.9	+10.4	+10.9	-0.9	-16.8	-0.0	-5.3	-0.2
NH ₄ ⁺	212.3	-13.1	-19.5	+2.0	-2.0	-3.8	-3.4	-4.8
CH ₃ NH ₃ ⁺	223.3	-15.4	-19.2	+1.2	-0.4	-3.8	-2.0	-4.0
1-aminobutaneH ⁺	228.2	-15.3	-18.6	+1.6	-0.1	-3.5	-1.2	-3.3
MUE		13.6	15.9	2.3	3.6	2.1	2.5	2.3
MSE		-11.3	-13.5	+2.1	-0.7	-1.9	-0.3	-1.8
MAX		17.1	19.5	5.3	16.8	3.8	5.3	4.8

^aThe molecules are given in the protonated form. The proton affinity is computed with the potential energies at 0 K without any zero-point energy correction. For the DFTB methods, the deviation is given as the difference of the G3B3 method ($E^{\text{method}} - E^{\text{G3B3}}$). The compilation of the molecules is taken from ref 39. ^bFor explanations, see Tables 1 and 2.

which is an important measure for the accuracy of proton transfer energetics between different donor and acceptor species.

We also benchmark DFT methods with medium-sized basis sets for proton affinities. The MUE of PBE/6-31+G(d,p) and B3LYP/6-31+G(d,p) in comparison to G3B3 is 4.7 and 2.5 kcal/mol, which is comparable to the performance of DFTB3. Note that the use of diffuse functions is essential here, and errors for calculations without diffuse functions are much larger (for details, see the Supporting Information). For example, the use of HF/4-31G for the description of a proton transfer reaction may not yield a correct description of the dynamics due to errors in the PAs of the donor and acceptor.⁸⁰

Overall, one can find a clear difference in the performance of DFTB2 and DFTB3 due to the inclusion of the third order terms, whereas DFTB3-dia and DFTB3 perform very similar on proton affinities. As for the binding energies, Yang et al.³⁹ compiled larger test sets and showed that DFTB3-dia overall improves the description of proton affinities. This is true for both, using calculated Hubbard derivatives or fitted Hubbard derivatives. With these findings, we also expect similar behavior for DFTB3. We have seen that an improved performance for both, hydrogen binding energies and proton affinities of DFTB2, is only found when including both extensions, the γ^h function and third order terms. Therefore, further benchmark tests are shown in the following for the combination of these extensions, and also the improvement of DFTB3 over DFTB3-dia will be discussed.

4.4. Proton Transfer Barriers. For testing proton transfer barriers, several simple models are considered. For the O...H...O models, we place a proton between two water molecules and between two hydroxide anions. The barriers are calculated for a fixed oxygen–oxygen distance with the shared proton at half the distance between both oxygens. All other hydrogen atoms are geometry optimized. For the relaxed structure, the shared proton is allowed to relax. While for the cationic complex the barriers calculated with MP2⁸¹/G3large⁶² are already well reproduced with DFTB2, large errors occur for the anionic model for large O–O distance. These errors are

completely removed for both DFTB3-dia and DFTB3. Table 9 summarizes the results, from which we note that the DFTB3 results represent a notable improvement over popular DFT methods with an intermediate basis set.

Similarly, proton transfer barriers for nitrogen species are tested. DFTB2 underestimates the barriers severely, while DFTB3/calc reduces this error and even slightly overestimates the barrier for the negatively charged complex. The DFTB models with fitted parameters show further improved results.

The proton transfer barriers for the models containing one oxygen and one nitrogen atom are computed keeping both heavy atoms fixed and translating the shared proton along the straight line between oxygen and nitrogen. The barrier is then given by the highest energy surrounded by two minima. For the relaxed structure, the shared proton is again geometry optimized together with all other hydrogen atoms. Rather large deviations are found for DFTB2 which are reduced with DFTB3-dia/calc and DFTB3/calc. Again, an overall good performance is found for the DFTB3-dia/fit and DFTB3/fit versions; the largest errors appear for $[\text{NH}_3\text{--H--H}_2\text{O}]^+$, where surprisingly DFT-GGA methods also reveal comparably large errors (see Table 9) in comparison to MP2/G3large.

The use of NHmod has the following consequences on barriers. The N–H bond is energetically shifted by about 10 kcal/mol, being more attractive in the binding region. The strength of the bond decreases with larger N–H distances. As a consequence, no barrier can be found for the models containing one oxygen and one nitrogen with small N–O distances. Here, we see that NHmod is not parametrized and not applicable to proton transfer barriers. Nevertheless, NHmod is a practical solution for correcting errors for proton affinities, as has been shown in several applications, e.g., ref 44. For models with two nitrogen atoms, we find very similar results for NHorg and NHmod. Future work will have to be concentrated on solving the hybridization problem and balancing N–H and O–H repulsive potentials such that proton transfer barriers with oxygen and nitrogen participation are described correctly.

Table 9. Proton Transfer Barrier in Kilocalories per Mole for a Fixed Distance (rXY) between the Heavy Atoms (X,Y ∈ {O,N}): Deviation of DFTB and DFT in Comparison to MP2/G3large^a

barrier	rXY	MP2	DFTB2	DFTB3-dia ^b		DFTB3 ^b		PBE ^c	B3LYP ^c
				calc	fit	calc	fit		
[H ₂ O-H-H ₂ O] ⁺	2.5	0.6	-0.6	-0.6	-0.6	-0.6	-0.6	-0.5	-0.4
	2.6	2.4	-1.4	-0.2	+0.0	-1.1	-0.8	-1.9	-1.0
	2.7	5.2	-1.1	+0.5	+0.8	-0.7	-0.3	-3.2	-1.5
	2.8	8.9	-1.3	+0.5	+0.9	-0.9	-0.4	-4.4	-1.9
[OH-H-OH] ⁻	2.5	0.5	-0.5	-0.5	-0.5	-0.5	-0.5	-0.5	-0.2
	2.6	2.3	-2.3	-0.6	-0.2	-0.7	+0.0	-1.6	-0.6
	2.7	5.2	-4.6	-0.0	+0.5	-0.2	+0.9	-2.7	-0.9
	2.8	8.8	-6.7	-0.0	+0.7	-0.3	+1.2	-3.7	-1.2
[NH ₃ -H-NH ₃] ⁺	2.6	0.4	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	-0.3
	2.7	1.9	-1.8	-1.0	-1.4	-1.7	-1.7	-1.9	-1.0
	2.8	4.4	-2.4	-0.5	-1.4	-2.0	-2.2	-3.0	-1.6
	2.9	7.7	-2.5	-0.1	-1.2	-2.1	-2.2	-4.1	-2.2
[NH ₂ -H-NH ₂] ⁻	2.5	0.1	-0.0	+0.3	-0.1	-0.1	-0.1	+0.0	+0.0
	2.6	1.4	-1.4	+5.2	-1.1	-0.4	-1.0	-1.2	-0.5
	2.7	3.5	-3.5	+5.5	-0.6	+1.5	-0.5	-2.1	-0.8
	2.8	6.3	-4.9	+6.9	+0.6	+2.9	+0.9	-3.0	-1.1
[NH ₃ -H-H ₂ O] ^{+d}	2.9	25.3	-8.2	-5.2	-4.2	-5.8	-5.3	- ^e	-2.7
	3.0	30.0	-9.5	-6.3	-5.4	-7.3	-6.7	-6.8	-3.3
	3.1	35.1	-11.0	-7.7	-6.8	-8.9	-8.2	-7.8	-3.8
	3.2	40.5	-12.5	-9.1	-8.2	-10.5	-9.8	-8.8	-4.2
[H ₂ O-H-NH ₃] ^{+d}	2.9	0.8	+0.7	+2.2	+1.1	+0.2	+0.1	- ^e	-0.8
	3.0	3.3	+0.9	+2.9	+1.5	+0.2	+0.1	-3.0	-1.6
	3.1	6.7	+0.3	+2.3	+0.9	-0.6	-0.7	-4.5	-2.3
	3.2	10.7	-0.8	+1.3	-0.2	-1.9	-2.0	-5.9	-2.9
[NH ₂ -H-OH] ^{-d}	2.8	10.1	-6.3	-4.6	+2.4	-2.5	+2.9	-3.3	-0.8
	2.9	14.2	-8.6	-5.2	+2.1	-3.1	+2.7	-4.2	-1.1
	3.0	18.6	-11.1	-6.0	+1.4	-3.9	+2.1	-5.1	-1.4
	3.1	23.3	-13.4	-6.7	+0.7	-4.7	+1.5	-5.9	-1.6
[OH-H-NH ₂] ^{-d}	2.8	4.4	-4.1	+14.4	+0.3	+6.6	+0.7	-2.8	-1.2
	2.9	7.8	-5.2	+15.8	+1.1	+7.9	+1.6	-4.0	-1.7
	3.0	11.6	-6.6	+16.5	+1.2	+8.4	+1.7	-4.9	-2.0
	3.1	15.8	-8.8	+16.2	+0.1	+7.8	+0.7	-5.8	-2.3

^a Barriers are computed as described in the text at 0 K, and no zero-point energy correction has been included. For the DFT and DFTB methods, the deviation is given as the difference of the MP2 method ($E^{\text{method}} - E^{\text{MP2}}$). For all models, the NHorg parameter set is applied. ^b For explanations, see Tables 1 and 2. ^c Basis set 6-31+G(d,p). ^d Barrier in comparison to the relaxed structure with the proton binding to the heavy atom that is written on the left-hand side of that proton. ^e A barrier does not exist.

To point this out more clearly, we take a look at [NH₃-H-H₂O]⁺, a model for a proton transfer between an amino acid with an acidic nitrogen in the side chain (lysine, histidine, arginine) and an oxygen. In the model, the nitrogen is sp³ hybridized, as would be the case for lysine, and we find an error of about 10 kcal/mol. Therefore, NHmod should be used. However, when doing so, the barrier vanishes; i.e., the energy monotonically rises as the hydrogen moves toward oxygen. Thus, proton affinities and proton transfer barriers can be well described for systems including histidine or arginine (applying NHorg), but special care must be taken for lysine. When NHmod is used, the proton affinity is described well, but not the barrier height of a proton transfer. The same problem arises for DNA proton transfer reactions, where the MUE is about 5 kcal/mol for proton affinities of DNA bases, as found for DFTB3-dia in ref 39.

4.5. Phosphorus-Containing Molecules. *4.5.1. Proton Affinity and Hydrolysis Energetics.* For phosphorus-containing

molecules, we first apply two tests from ref 40. Table 10 shows 18 proton affinities of biological relevance; in Table 11, elementary steps for a representative set of phosphate hydrolysis reactions are listed, which include the hydrolysis of monophosphate ester (MMP) and dimethyl monophosphate ester (DMP) with different protonation states, numbers of water involved, and dissociative/associative mechanisms.

Proton affinities are generally overestimated for DFTB2 and DFTB3/calc. The MUE shows even worse results for DFTB3/calc than for DFTB2; however, similar to the case for the proton affinities with acidic oxygen, the deviation in comparison to the MP2 results is much more consistent with DFTB3. For DFTB3, all proton affinities are overestimated, the MSE being as large as the MUE (12.6 kcal/mol); additionally, the maximal absolute deviation is smaller than that for DFTB2. As discussed above, about 6 kcal/mol of this error is due to the overbinding of the O-H bond, which can be removed by refitting the O-H

Table 10. 18 Proton Affinities for Phosphorous Containing Molecules in kcal/mol: Deviation of DFTB in Comparison to G3B3^a

molecule ^b	G3B3	DFTB2	DFTB3-dia ^c		DFTB3 ^c	
			calc	fit	calc	fit
H ₃ PO ₄	334.0	+17.1	+23.9	+18.5	+18.3	+5.5
H ₂ PO ₂ ⁻	464.5	+26.2	+26.6	+20.0	+17.2	-4.3
DMPH ^d	336.3	+9.6	+20.3	+14.9	+15.9	+4.8
MMP ^d	336.7	+12.0	+20.9	+15.4	+15.8	+3.8
MMP ^{-d}	460.5	+21.5	+26.3	+19.9	+18.3	-1.2
PH ₃ OH ⁺	201.6	-8.6	+7.1	+2.7	+4.8	-0.0
PH ₂ OHOH ⁺	201.6	-2.8	+10.8	+6.3	+8.2	+2.1
PHOHOHOH ⁺	200.8	+4.5	+16.3	+11.7	+13.6	+6.2
PH ₂ (OH)=O	336.6	+3.0	+15.7	+10.4	+12.2	+3.3
PH(OH)(OH)=O	334.7	+10.7	+20.4	+15.0	+16.0	+5.3
P(O)(OH)(-O-CH ₂ CH ₂ -O-)	336.3	+7.2	+17.7	+12.3	+13.4	+2.4
P(OH)(OH)(-O-CH ₂ CH ₂ -O-)(OH*)	359.0	-3.3	+18.6	+12.8	+12.9	+0.3
P(OH*)(OH)(-O-CH ₂ CH ₂ -O-)(OH)	350.4	+6.7	+15.9	+10.6	+11.0	-0.4
P(OH*)(OH)(-O-CH ₂ CH ₂ -O-)(OCH ₃)	351.2	+1.8	+12.8	+7.4	+8.9	-1.4
P(OH)(OCH ₃)(-O-CH ₂ CH ₂ -O-)(OH*)	359.6	-8.3	+7.9	+2.1	+3.3	-7.2
P(OH*)(OCH ₃)(-O-CH ₂ CH ₂ -O-)(OH)	352.9	+3.6	+14.5	+9.2	+10.1	-0.5
P(OH)(OH)(OH)(OH*)(OH)_ax	357.3	+4.0	+21.8	+15.9	+14.2	-1.2
P(OH)(OH)(OH)(OH*)(OH)_eq ^e	347.0	+14.0				-0.0
MUE		9.2	17.5	12.1	12.6	2.8
MSE		+6.6	+17.5	+12.1	+12.6	+1.0
MAX		26.2	26.6	20.0	18.3	7.2

^a The proton affinity is computed with the potential energies at 0 K without any zero-point energy correction. For the DFTB methods, the deviation is given as the difference of the G3B3 method ($E^{\text{method}} - E^{\text{G3B3}}$). The compilation of the molecules is taken from ref 40. ^b The molecules are given in the protonated form. ^c For explanations, see Tables 1 and 2. ^d "DMPH" refers to dimethyl hydrogen phosphate, "MMP" to P(O)(OH)(OH)(OCH₃), and "MMP⁻" to P(O)(O)(OH)(OCH₃)⁻. ^e The molecule P(OH)(OH)(OH)(OH*)(OH)_eq dissociates, forming H₂O for DFTB3-dia^c and DFTB3/calc. Depending on the basis set, this dissociation also occurs for the DFT functionals PBE and B3LYP, e.g., dissociation for basis set 6-311G(2d,2p), no dissociation for basis set cc-pVTZ.

repulsive potential. The remaining error may be reduced by fitting the P Hubbard derivative. For now, the method of choice is DFTB3/fit, for which the MUE is only 2.8 kcal/mol. We want to point out that the Hubbard derivatives for C, H, N, and O are taken from the fit on nonphosphate molecules, and only U_{P}^{d} is fitted to the 18 listed proton affinities; this is in contrast to earlier work, where the best performance for DFTB3-dia^c could only be achieved by fitting *all* parameters at once.

Table 10 also shows the results for DFTB3-dia^c/calc, which look similar to DFTB3/calc. However, when using DFTB3-dia^c/fit, the error cannot be reduced as much as is the case for DFTB3/fit. We find that the parameter U_{P}^{d} has very small influence on the proton affinities. The MUE ranges from 10.7 to 12.5 kcal/mol when choosing U_{P}^{d} in the range of -0.40 to -0.04 atomic units; therefore, we keep the calculated parameter $U_{\text{P}}^{\text{d}} = -0.07$ H. This observation highlights that DFTB3-dia^c/fit does not properly account for some part of the interactions within these molecules; i.e., the flexibility of the model is not sufficient to yield good results for nonphosphate and phosphate molecules at the same time.

For the hydrolysis reactions, the MUE for DFTB2 is 4.4 kcal/mol and is only slightly reduced for DFTB3/fit (Table 11). Note that for the latter, the parameter U_{P}^{d} is fitted to the proton affinities only. A special fit also for these reactions does not improve this situation significantly. Surprisingly, DFTB3-dia^c performs somehow in a superior way with a MUE of 3.2 kcal/mol.

4.5.2. Additional Discussion of Transferability of Parameters. Earlier extensions of DFTB2 have suggested a lack of general

transferability; for example, the two phosphorus related parameter sets (see additional discussions in the next subsection), SCC-DFTBPA and SCC-DFTBPR,⁴⁰ need to be developed for different properties. Both sets are based on DFTB3-dia^c (without the γ^{h} function) with fitted Hubbard derivatives and an additional empirical Gaussian term (with three additional parameters) to adjust the Hubbard derivatives within the SCC procedure. SCC-DFTBPA is specifically designed for proton affinities of phosphorus-containing molecules and yields a MUE for the 18 proton affinities of Table 10 of only 2.6 kcal/mol but performs in an inferior way for proton affinities of nonphosphate molecules. SCC-DFTBPR, on the other hand, is designed for the hydrolysis reactions of Table 11 and shows a MUE for these reactions of only 2.4 kcal/mol but is less accurate for the proton affinities (in particular for nonphosphate molecules).

DFTB3 is a consistent extension of our model and transferable to a wide range of chemical properties. Instead of different methods with a different number of parameters (six or nine parameters additionally to the ones from DFTB2) we now have a method at hand that shows an overall good performance for binding energies and proton affinities of nonphosphate and phosphate molecules using only six additional parameters in comparison to DFTB2 (ζ , U_{C}^{d} , U_{H}^{d} , U_{N}^{d} , U_{O}^{d} , U_{P}^{d}). A limitation is found, however, for the hydrolysis reactions, for which only a slight improvement is achieved in comparison to DFTB2. DFTB3 is not performing as well as SCC-DFTBPR in that respect, which suggests that further improvements are

Table 11. Deviation of Exothermicity and Barrier Height from the DFTB Variants in Comparison to MP2/G3Large Single Point Calculations at B3LYP/6-31+G(d,p) Structures for 37 Elementary Steps in the Hydrolysis of MMP and DMP^a

process ^b	MP2	DFTB2	DFTB3-dia ^c		DFTB3 ^c	
			calc	fit	calc	fit
com1 → ts1 (MMP,B)	31.0	-1.0	-3.9	-4.2	-6.7	-7.2
com1 → int1 (MMP,E)	30.6	-2.2	-2.4	-2.6	-6.1	-7.0
com1 → ts1_2 (MMP,B)	41.5	+0.8	-1.7	-1.9	-3.9	-3.3
com1 → int1_2 (MMP,E)	31.0	-4.4	+0.5	+0.3	-4.6	-5.9
int1_2 → ts2_0 (MMP,B)	11.9	-3.0	-3.0	-3.0	-0.2	+2.7
int1_2 → ts2 (MMP,B)	3.6	-5.5	-4.9	-4.9	-2.8	-1.1
int1_2 → com2 (MMP,E)	-28.8	+2.1	+0.0	+0.1	+3.3	+4.4
com1 → diss_tsa (MMP,B)	36.8	+4.6	+2.2	+2.4	+3.3	+4.9
com1 → diss_int (MMP,E)	19.6	-7.3	-7.3	-6.7	-3.2	+0.3
com1_w2 → ts1_2_w2 (MMP,B)	39.9	-8.3	-12.3	-12.1	-13.6	-12.7
com1_w2 → int1_2a_w2 (MMP,E)	28.0	-5.0	-1.5	-1.8	-6.6	-7.9
int1_2a_w2 → int1_2_w2 (MMP,E)	0.4	+0.3	+1.7	+1.5	+2.5	+2.5
int1_2_w2 → ts2_0_w2 (MMP,B)	11.4	-4.1	-10.8	-10.5	-7.7	-5.4
com1_da → ts1_da (MMP,B)	55.0	-22.4	-8.8	-9.1	-7.1	-0.2
com1_da → int_da (MMP,E)	4.5	-3.3	+0.6	+0.5	-1.6	-2.0
com1 → ts1 (DMP,B)	38.6	-1.7	-5.3	-5.8	-7.1	-7.3
com1 → int1 (DMP,E)	35.4	-5.7	-2.2	-2.4	-6.4	-7.6
int1 → int1_2 (DMP,E)	1.3	-3.2	-0.3	-0.3	+0.4	+1.3
int1_2 → ts2 (DMP,B)	0.6	+0.2	-0.5	-0.6	+0.4	+1.2
int1_2 → com2 (DMP,E)	-35.2	+7.0	+4.0	+4.1	+5.9	+6.0
n_com1 → n_ts3 (DMP,B)	33.6	+4.9	+3.0	+2.7	+0.7	+0.2
n_com1 → n_int1 (DMP,E)	13.2	-3.7	-0.0	-0.4	-3.8	-4.9
n_int1 → n_ts4 (DMP,B)	22.9	+6.4	+4.1	+4.0	+4.8	+5.1
n_int1 → n_com2 (DMP,E)	-15.8	+2.2	-0.1	+0.0	+3.2	+3.7
DMP_P → diss_ts (DMP,B)	40.9	+11.6	+8.3	+8.1	+8.8	+8.8
DMP_P → diss_prod (DMP,E)	28.2	+0.6	-1.7	-1.8	-0.0	-0.9
diss_prod2 → diss_ts2 (DMP,B)	13.5	+13.2	+11.3	+11.2	+9.8	+10.4
diss_prod2 → MMP_P (DMP,E)	-29.8	+0.8	+2.9	+2.8	+0.3	+0.6
diss_w_reac → diss_w_ts (DMP,B)	20.9	+5.9	+2.2	+2.2	+2.3	+0.7
diss_w_reac → diss_w_prod (DMP,E)	18.4	+4.3	+0.2	+0.1	+1.4	+0.3
diss_w_prod2 → diss_w_ts2 (DMP,B)	1.9	+2.7	+2.2	+2.2	+0.9	+0.3
diss_w_prod2 → diss_w_reac2 (DMP,E)	-21.0	-2.5	+0.4	+0.4	-1.4	-0.7
n_w_com1 → n_w_ts3 (DMP,B)	28.2	-2.7	-4.4	-4.4	-7.6	-8.9
n_w_com1 → n_w_int1 (DMP,E)	13.1	-4.1	-0.7	-1.0	-4.5	-5.8
n_w_int1 → n_w_int2 (DMP,E)	-0.5	+0.1	+0.7	+0.7	+0.7	+0.6
n_w_int2 → n_w_ts4 (DMP,B)	15.1	+2.0	-2.2	-2.0	-2.3	-3.1
n_w_int2 → n_w_com2 (DMP,E)	-13.0	+1.4	-0.7	-0.7	+3.0	+4.1
MUE		4.4	3.2	3.2	4.0	4.1
MSE		-0.5	-0.8	-0.9	-1.2	-0.9
MAX		22.4	12.3	12.1	13.6	12.7

^a Compilation from ref 40. No zero-point corrections are included in either exothermicity or barrier heights. All quantities are given in kcal/mol. ^b The processes are labeled as in ref 40. "E" stands for "Exothermicity", "B" for "Barrier". All structures are listed in the Supporting Information. ^c For explanations, see Table 2.

necessary for the phosphorus parameters and/or for the DFTB formalism.

4.5.3. *Geometry and Nonisodesmic Reactions.* In several publications, phosphorous parameters (electronic and repulsive parameters) for DFTB2 have been used,^{40,42,45,47,82,83} and the parametrization procedure has been described in ref 40.

Geometrical properties are tested on 35 molecules in the gas phase, including phosphorus-containing acids in different

protonation states. All DFTB versions perform quite well in comparison to B3LYP/6-31+G(d,p), but there are specific bond types that show a general trend of being too short or too long, as summarized in Table 12. For example, bonds for the pairs P–P, P–S, and P=S are typically too long. A significant difference between DFTB2 and DFTB3 can be found for the O–P single bond length, which is rather too short for the latter in comparison to B3LYP/6-31+G(d,p). This can be most dramatically seen in

Table 12. DFTB2 and DFTB3 Errors for the Bond Lengths of 35 Phosphorus-Containing Molecules^a in Comparison to B3LYP/6-31+G(d,p)

bond type ^b	n ^c	DFTB2			DFTB3/calc			DFTB3/fit ^d		
		MAX	MSE	MUE	MAX	MSE	MUE	MAX	MSE	MUE
rC–P	6	0.042	+0.022	0.022	0.041	+0.018	0.018	0.041	+0.021	0.021
rC=P	1	0.056	+0.056	0.056	0.049	+0.049	0.049			
rH–P	9	0.061	+0.024	0.024	0.047	+0.020	0.021	0.044	+0.018	0.019
rN–P	1	0.001	+0.001	0.001	0.001	+0.001	0.001	0.001	+0.001	0.001
rO–P	43	0.404	+0.001	0.029	0.150	−0.022	0.026	0.200	−0.029	0.031
rO=P	33	0.029	+0.007	0.011	0.027	+0.008	0.011	0.017	+0.003	0.006
rP–P	4	0.109	+0.077	0.077	0.118	+0.081	0.081	0.148	+0.098	0.098
rP#P	1	0.003	+0.003	0.003	0.003	+0.003	0.003	0.003	+0.003	0.003
rP–S	5	0.164	+0.121	0.121	0.137	+0.103	0.103	0.128	+0.098	0.098
rP=S	3	0.070	+0.062	0.062	0.064	+0.059	0.059	0.061	+0.059	0.059
rOHhb	3	0.141	−0.138	0.138	0.225	−0.222	0.222	0.192	−0.186	0.186
overall ^e	196	0.404	+0.010	0.022	0.225	+0.002	0.021	0.200	−0.000	0.021

^a Geometries for all molecules are listed in the Supporting Information. ^b Bond situations between two atom types, “–” means a single bond, “=” a double bond, rP#P is the bond length of the molecule P₂, and rOHhb represents hydrogen bonds between a phosphate group and a water. All specifications can be found in the Supporting Information. ^c Number of comparisons. ^d Trimethylmethylenephosphorane does not converge and is excluded from the statistics. ^e Also the bond types rCC, rCH, rCO, rC=O, rHO, rCS, and rHS are included in the overall performance.

the example where an acetate is linked via one oxygen atom to a phosphate group ([CH₃COO–PO₃]^{2–}). For DFTB2, this molecule almost dissociates with an O–P distance of 2.372 Å; for DFTB3, it is too short (1.818 Å) in comparison to the B3LYP result (1.968 Å). As already shown in ref 40, DFTB2 has hydrogen bonding distances that are too small for water–phosphate bonds. This problem is not resolved for DFTB3 and needs to be addressed with an improved description of Pauli repulsion in the DFTB2 framework. The MUE for bond angles is between 2.5° and 3.0° for all DFTB versions. Further details can be found in the Supporting Information.

For an additional test of chemical reactions beyond hydrolysis, 10 reactions have been carried out in which the bonding situation changes; e.g., a O–P bond is exchanged to a H–P bond (nine reactions are hydrogenations). We find large deviations, the MUE being around 50 kcal/mol for all DFTB versions; details can be found in the Supporting Information. Among other shortcomings, the most important one seems to be the eminent shortbinding of the PO bonds. Thus, while the current phosphorus parameters might well be used for geometries (while special care is necessary for some bond types, see above), proton affinities and hydrolysis reactions, it should not be applied for nonisodesmic reactions.

5. CONCLUSION

We have presented DFTB3, a new method that extends the standard second order DFTB2 (formally SCC-DFTB) by two conceptually independent improvements. DFTB3 maintains the strengths of DFTB2, such as rapid computation of large scale molecular systems with reliable geometry, but improves transferability and overall accuracy for several properties.

The first concept is the γ^h function ameliorating the electron–electron interaction of charge fluctuation. The γ^h function corrects the original function, which incorrectly imposes a linear relation between the chemical hardness and the atomic size. This relationship is only valid within one row of the periodic table and particularly fails when interactions of first row atoms with

hydrogen are involved. We therefore introduced one additional, purely empirical parameter (ζ), which can be adjusted to a single reference system like the water dimer. Previous tests have shown that this improves the performance of DFTB2 for hydrogen bonding systematically. Therefore, this correction does not introduce additional terms to total energy in an *ad hoc* fashion but establishes a consistent improvement of the electron–electron interaction in the second (and third) order terms of DFTB2 (DFTB3). As a result, the mean unsigned error for hydrogen bonding energies drops from 8.0 kcal/mol for DFTB2 to 4.0 kcal/mol for DFTB2- γ^h for our fairly broad sets of test systems. A drawback is found for the hydrogen bond lengths which turn out to be too short. [In principle, these repulsive energy potentials are intimately coupled to the electronic DFTB terms with which they have been determined. Therefore, they have to be refitted when the DFTB Hamiltonian is modified.]

The second improvement concerns the extension of DFTB2 to include third order terms of the Taylor series expansion of the DFT exchange–correlation energy. The third order terms cause the chemical hardness (Hubbard parameter) of an atom to be dependent on its charge, which becomes particularly important for the description of systems with localized charges. One additional parameter is introduced for each element, the Hubbard derivative with respect to charge, which can be either computed from DFT for atoms or can be fitted. With the first approach, the DFTB3- γ method does not involve any new empirical parameters. Geometries for charged molecules are slightly improved. Regarding proton affinities, the errors become consistently overestimated in contrast to an underestimation for negatively charged systems and an overestimation for positively charged systems with DFTB2.

The combination of both improvements in DFTB3 also combines the effects. The accuracy of DFTB2 for geometries of C-, H-, N-, and O-containing molecules is maintained. For charged molecules, a slight geometrical improvement is found, whereas hydrogen bonds are consistently too short. The mean unsigned error for our set of hydrogen binding energies drops below 3.0 kcal/mol. It should be noted that this improved DFTB

model outperforms standard DFT functionals using medium-sized basis sets without correction for BSSE, a methodology typically used in, for example, QM/MM applications to biological systems. [Note that adding empirical dispersion corrections to DFT-GGA would even worsen the situation, since DFT overbinds the H-bonded complexes already due to BSSE.] For proton affinities, the mean unsigned error is not significantly reduced when using calculated Hubbard derivatives. However, we have shown that the remaining errors do not arise due to third order approximations (and neither the γ^h function) but result from the repulsive potential terms of second order DFTB2; i.e., the errors could be removed in principle by reoptimizing the DFTB repulsive potentials, with which an empirical fitting of the Hubbard derivative parameters is likely no longer necessary. For the time being, we also present empirically fitted parameters (Hubbard derivatives and ζ), which result in a significant improvement over DFTB2. The mean unsigned deviation for our oxygen-containing test systems in comparison to G3B3 results are 5.8, 2.9, and 2.5 kcal/mol for DFTB2, DFTB3 with fitted ζ and Hubbard derivatives, and B3LYP/6-31+G(d,p), respectively.

We have also shown that the energy of a proton is a constant and not equal to zero for DFTB2 (and DFTB3) due to the neglect of atomic contributions within the repulsive energy contribution. There are different eligible ways of how to compute this constant, leading to different constants in order to obtain an absolute proton affinity. For reasons of consistency, we used the constant as calculated directly from the respective level of theory (DFTB2 or DFTB3). We emphasize that for most applications only relative proton affinities are important; i.e., the value of this constant does not matter at all. Only for specific applications where the absolute proton affinity is needed does the value of that constant become important, e.g., determining the pK_a of a molecule. An empirical but helpful choice different than fitting parameters would then be to use a constant which compensates the consistent over- or underestimation of the respective DFTB variant.

In earlier work, we have already implemented and tested the diagonal part of the third order corrections and provided different parametrizations (Hubbard derivatives, ζ , and in some cases also additional parameters).^{39,40} In our comparison of DFTB3-diag and DFTB3, the newly implemented off-diagonal terms do not seem to lead to a large improvement for molecules consisting of O, N, C, and H, except for the NH_2 molecule, since the diagonal part is already quite accurate. The most significant advantage of DFTB3 over DFTB3-diag and earlier published extensions of DFTB2 is its consistent performance for hydrogen bonding energies and proton affinities including atoms of type C, H, N, O, and P. While all earlier extensions needed different parametrizations for different properties, DFTB3 with fitted ζ and Hubbard derivatives is more transferable and covers all properties with a single parametrization. One persistent limitation is found for phosphate hydrolysis reactions, where a model based on DFTB3-diag with an empirical Gaussian term and “reaction specific” parametrization of the Hubbard derivatives (SCC-DFTBPR)⁴⁰ is still needed for better accuracy.

Despite all progress, major limitations for DFTB3 remain. First, the error of proton affinities of nitrogen-containing molecules seems to correlate with the hybridization state of nitrogen. We discussed the use of different repulsive potentials, NHorg and NHmix, which provides a pragmatic way for calculating accurate proton affinities but is unreliable for studying reactions and

proton transfer barriers. Moreover, the scheme is conceptually unsatisfactory. Second, the hydrogen bond lengths are generally too short, and third, large errors are found for nonisodesmic reactions of phosphorus-containing species. Addressing these limitations requires developing new electronic and repulsive parameters (or formulations) for DFTB3.

■ ASSOCIATED CONTENT

S Supporting Information. Detailed derivation of the Kohn–Sham equations, atomic forces, and the γ^h function within the DFTB3 formalism (PDF). Excel file for molecular geometries and computed data. This information is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: marcus.elstner@kit.edu.

■ ACKNOWLEDGMENT

We thank Dr. Yang Yang and Puja Goyal for providing us compilations and geometries of their earlier works and Dr. Bálint Aradi for implementing the new methodologies into DFTB+ and for helpful discussions. This work was partially supported by NIH grant R01GM084028 (to Q.C.).

■ REFERENCES

- (1) Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4899.
- (2) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
- (3) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 209.
- (4) Rocha, G. B.; Freire, R. O.; Simas, A. M.; Stewart, J. J. P. *J. Comput. Chem.* **2006**, *27*, 1101.
- (5) Stewart, J. J. P. *J. Mol. Model.* **2007**, *13*, 1173.
- (6) Repasky, M. P.; Chandrasekhar, J.; Jorgensen, W. L. *J. Comput. Chem.* **2002**, *23*, 1601.
- (7) Kolb, M.; Thiel, W. *J. Comput. Chem.* **1993**, *14*, 775.
- (8) Weber, W.; Thiel, W. *Theor. Chem. Acc.* **2000**, *103*, 495.
- (9) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260.
- (10) Elstner, M.; Frauenheim, T.; McKelvey, J.; Seifert, G. *J. Phys. Chem. A* **2007**, *111*, 5607.
- (11) Porezag, D.; Frauenheim, T.; Köhler, T.; Seifert, G.; Kaschner, R. *Phys. Rev. B* **1995**, *51*, 12947.
- (12) Seifert, G.; Porezag, D.; Frauenheim, T. *Int. J. Quantum Chem.* **1996**, *58*, 185.
- (13) Seifert, G.; Eschrig, H. *Phys. Stat. Sol. B* **1985**, *127*, 573.
- (14) Seifert, G.; Eschrig, H.; Bieger, W. Z. *Phys. Chem. (Leipzig)* **1986**, *267*, 529.
- (15) Foulkes, W. M. C.; Haydock, R. *Phys. Rev. B* **1989**, *39*, 12520.
- (16) Harris, J. *Phys. Rev. B* **1985**, *31*, 1770.
- (17) Hazebroucq, S.; Picard, G. S.; Adamo, C.; Heine, T.; Gemming, S.; Seifert, G. *J. Chem. Phys.* **2005**, *123*, 1.
- (18) Frauenheim, T.; Seifert, G.; Elstner, M.; Hajnal, Z.; Jungnickel, G.; Porezag, D.; Suhai, S.; Scholz, R. *Phys. Stat. Sol. B* **2000**, *217*, 41.
- (19) Elstner, M.; Frauenheim, T.; Kaxiras, E.; Seifert, G.; Suhai, S. *Phys. Stat. Sol. B* **2000**, *217*, 357.
- (20) Frauenheim, T.; Seifert, G.; Elstner, M.; Niehaus, T.; Köhler, C.; Amkreutz, M.; Sternberg, M.; Hajnal, Z.; Di Carlo, A.; Suhai, S. *J. Phys.: Cond. Matter* **2002**, *14*, 3015.
- (21) Elstner, M.; Frauenheim, T.; Suhai, S. *THEOCHEM* **2003**, *632*, 29.
- (22) Elstner, M. *Theor. Chem. Acc.* **2006**, *116*, 316.

- (23) Riccardi, D.; Schaefer, P.; Yang, Y.; Yu, H.; Ghosh, N.; Prat-Resina, X.; König, P.; Li, G.; Xu, D.; Guo, H.; Elstner, M.; Cui, Q. *J. Phys. Chem. B* **2006**, *110*, 6458.
- (24) Density Functional Tight Binding: Contributions from the American Chemical Society Symposium: *J. Phys. Chem. A*, **2007**, *111*, 5607.
- (25) Sattelmeyer, K. W.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. A* **2006**, *110*, 13551.
- (26) Otte, N.; Scholten, M.; Thiel, W. *J. Phys. Chem. A* **2007**, *111*, 5751.
- (27) Gaus, M.; Chou, C.; Witek, H.; Elstner, M. *J. Phys. Chem. A* **2009**, *113*, 11866.
- (28) Elstner, M.; Jalkanen, K. J.; Knapp-Mohammady, M.; Frauenheim, T.; Suhai, S. *Chem. Phys.* **2000**, *256*, 15.
- (29) Elstner, M.; Jalkanen, K. J.; Knapp-Mohammady, M.; Frauenheim, T.; Suhai, S. *Chem. Phys.* **2001**, *263*, 203.
- (30) Seabra, G. D. M.; Walker, R. C.; Roitberg, A. E. *J. Phys. Chem. A* **2009**, *113*, 11938.
- (31) Krüger, T.; Elstner, M.; Schifflers, P.; Frauenheim, T. *J. Chem. Phys.* **2005**, *122*, 1.
- (32) Witek, H. A.; Morokuma, K. *J. Comput. Chem.* **2004**, *25*, 1858.
- (33) Witek, H. A.; Morokuma, K.; Stradomska, A. *J. Chem. Phys.* **2004**, *121*, 5171.
- (34) Witek, H. A.; Morokuma, K.; Stradomska, A. *J. Theor. Comput. Chem.* **2005**, *4*, 639.
- (35) Witek, H. A.; Irle, S.; Zheng, G.; De Jong, W. A.; Morokuma, K. *J. Chem. Phys.* **2006**, *125*, 214706.
- (36) Elstner, M.; Hobza, P.; Frauenheim, T.; Suhai, S.; Kaxiras, E. *J. Chem. Phys.* **2001**, *114*, 5149.
- (37) Liu, H.; Elstner, M.; Kaxiras, E.; Frauenheim, T.; Hermans, J.; Yang, W. *Proteins: Struct, Funct., Genet.* **2001**, *44*, 484.
- (38) Elstner, M. *J. Phys. Chem. A* **2007**, *111*, 5614.
- (39) Yang, Y.; Yu, H.; York, D.; Cui, Q.; Elstner, M. *J. Phys. Chem. A* **2007**, *111*, 10861.
- (40) Yang, Y.; Yu, H.; York, D.; Elstner, M.; Cui, Q. *J. Chem. Theory Comput.* **2008**, *4*, 2067.
- (41) Riccardi, D.; Yang, S.; Cui, Q. *Biochim. Biophys. Acta* **2010**, *1804*, 342.
- (42) Yang, Y.; Cui, Q. *J. Phys. Chem. A* **2009**, *113*, 12439.
- (43) Phatak, P.; Frähmcke, J. S.; Wanko, M.; Hoffmann, M.; Strodel, P.; Smith, J. C.; Suhai, S.; Bondar, A.; Elstner, M. *J. Am. Chem. Soc.* **2009**, *131*, 7064.
- (44) Phatak, P.; Ghosh, N.; Yu, H.; Cui, Q.; Elstner, M. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 19672.
- (45) Yang, Y.; Yu, H.; Cui, Q. *J. Mol. Biol.* **2008**, *381*, 1407.
- (46) Riccardi, D.; König, P.; Guo, H.; Cui, Q. *Biochemistry* **2008**, *47*, 2369.
- (47) Yang, Y.; Cui, Q. *J. Phys. Chem. B* **2009**, *113*, 4930.
- (48) Bondar, A.; Fischer, S.; Smith, J. C.; Elstner, M.; Suhai, S. *J. Am. Chem. Soc.* **2004**, *126*, 14668.
- (49) Seifert, G. *J. Phys. Chem. A* **2007**, *111*, 5609.
- (50) Janak, J. F. *Phys. Rev. B* **1978**, *18*, 7165.
- (51) Politzer, P.; Murray, J. S.; Lane, P. *J. Comput. Chem.* **2003**, *24*, 505.
- (52) Winget, P.; Selçuki, C.; Horn, A. H. C.; Martin, B.; Clark, T. *Theor. Chem. Acc.* **2003**, *110*, 254.
- (53) Zhou, H.; Tajkhorshid, E.; Frauenheim, T.; Suhai, S.; Elstner, M. *Chem. Phys.* **2002**, *277*, 91.
- (54) Elstner, M.; Cui, Q.; Munih, P.; Kaxiras, E.; Frauenheim, T.; Karplus, M. *J. Comput. Chem.* **2003**, *24*, 565.
- (55) Ghosh, N.; Xavier, P.-R.; Gunner, M. R.; Cui, Q. *Biochemistry* **2009**, *48*, 2468.
- (56) Witek, H. A.; Köhler, C.; Frauenheim, T.; Morokuma, K.; Elstner, M. *J. Phys. Chem. A* **2007**, *111*, 5712.
- (57) Elstner, M. Ph.D. thesis, Universität-Gesamthochschule Paderborn, Soest, Germany, 1998.
- (58) Witek, H. A.; Irle, S.; Morokuma, K. *J. Chem. Phys.* **2004**, *121*, 5163.
- (59) Knaup, J. M.; Hourahine, B.; Frauenheim, T. *J. Phys. Chem. A* **2007**, *111*, 5637.
- (60) Niehaus, T. A.; Elstner, M.; Frauenheim, T.; Suhai, S. *THEO-CHEM* **2001**, *541*, 185.
- (61) Klopper, W.; Van Duijneveldt-van De Rijdt, J.; Van Duijneveldt, F. B. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2227.
- (62) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1998**, *109*, 7764.
- (63) Baboul, A. G.; Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **1999**, *110*, 7650.
- (64) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865.
- (65) 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.
- (66) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **1997**, *106*, 1063.
- (67) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
- (68) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
- (69) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623.
- (70) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007.
- (71) Xu, X.; Goddard, W. A., III. *J. Phys. Chem. A* **2004**, *108*, 2305.
- (72) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553.
- (73) Simon, S.; Duran, M.; Dannenberg, J. J. *J. Chem. Phys.* **1996**, *105*, 11024.
- (74) Ireta, J.; Neugebauer, J.; Scheffler, M. *J. Phys. Chem. A* **2004**, *108*, 5692.
- (75) Zhao, Y.; Truhlar, D. *J. Chem. Theory Comput.* **2005**, *1*, 415.
- (76) Santra, B.; Michaelides, A.; Scheffler, M. *J. Chem. Phys.* **2007**, *127*, 184104.
- (77) Rao, L.; Ke, H.; Fu, G.; Xu, X.; Yan, Y. *J. Chem. Theory Comput.* **2009**, *5*, 86.
- (78) Baer, M.; Mathias, G.; Kuo, I.-W.; Tobias, D. J.; Mundy, C. J.; Marx, D. *ChemPhysChem* **2008**, *9*, 2703.
- (79) Range, K.; Riccardi, D.; Cui, Q.; Elstner, M.; York, D. M. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3070.
- (80) Lee, Y.-S.; Krauss, M. *J. Am. Chem. Soc.* **2004**, *126*, 2225.
- (81) Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618.
- (82) Zhang, X.; Harrison, D. H. T.; Cui, Q. *J. Am. Chem. Soc.* **2002**, *124*, 14871.
- (83) Yang, Y.; Cui, Q. *J. Phys. Chem. B* **2007**, *111*, 3999.

Dynamic Precision for Electron Repulsion Integral Evaluation on Graphical Processing Units (GPUs)

Nathan Luehr, Ivan S. Ufimtsev, and Todd J. Martínez*

PULSE Institute and Department of Chemistry, Stanford University, Stanford, California 94305, United States
SLAC National Accelerator Laboratory, Menlo Park, California 94025, United States

S Supporting Information

ABSTRACT: It has recently been demonstrated that novel streaming architectures found in consumer video gaming hardware such as graphical processing units (GPUs) are well-suited to a broad range of computations including electronic structure theory (quantum chemistry). Although recent GPUs have developed robust support for double precision arithmetic, they continue to provide 2–8× more hardware units for single precision. In order to maximize performance on GPU architectures, we present a technique of dynamically selecting double or single precision evaluation for electron repulsion integrals (ERIs) in Hartree–Fock and density functional self-consistent field (SCF) calculations. We show that precision error can be effectively controlled by evaluating only the largest integrals in double precision. By dynamically scaling the precision cutoff over the course of the SCF procedure, we arrive at a scheme that minimizes the number of double precision integral evaluations for any desired accuracy. This dynamic precision scheme is shown to be effective for an array of molecules ranging in size from 20 to nearly 2000 atoms.

INTRODUCTION

It has recently been recognized that consumer video game hardware is well suited to many tasks in computational chemistry, including electronic structure theory,^{1–10} *ab initio* molecular dynamics,¹¹ and empirical force-field-based molecular dynamics.^{8,12–14} The emergence of the CUDA development framework from NVIDIA has made it much easier to repurpose this hardware for scientific computing,¹⁵ compared to early efforts on similar architectures that had to resort to low level instructions.¹⁶ Nevertheless, efficient use of graphical processing units (GPUs) requires careful attention to some specialized hardware constraints such as memory access patterns and non-uniform efficiency of floating point arithmetic in different precision. Furthermore, GPUs have been carefully designed for maximum performance in specific graphics processing tasks and are otherwise severely limited. It is unlikely that these limitations will be fully eliminated because in large part they provide the foundation of the GPUs computational prowess.

The first CUDA-enabled GPUs had no support for double precision arithmetic, demanding care in their use for quantum chemistry applications. The latest GPUs fully support double precision arithmetic, with stunning performance in the range of several hundred GFLOPS, well beyond that of traditional processors (CPUs). Nevertheless, single precision continues to maintain between 2× and 8× more instruction units than double precision on the latest generation of GPUs. This disparity stems from the hardware's pedigree in graphics, where there is little need for double precision accuracy, and the necessary increase in circuitry is difficult to justify. Single precision may exhibit further performance advantages as a result of its smaller memory footprint, which reduces data bandwidth requirements² and increases the number of values that can be cached in registers. Thus, for maximum

performance, it remains important to favor single precision arithmetic as much as possible on GPUs.

To balance GPU performance with chemical accuracy, quantum chemistry implementations have adopted mixed precision approaches in which double precision operations are added sparingly to an otherwise single precision calculation. Matrix multiplication in the context of resolution-of-the-identity Møller–Plesset perturbation theory has been shown to provide accurate mixed precision results, even when the majority of operations are carried out in single precision.^{4,6} Single precision ERI evaluation has been successfully augmented with double precision accumulation into the matrix elements of the Coulomb and exchange operators.^{3,5} “Double precision accumulation” simply means that the ERIs are evaluated in single precision, but a double precision variable is used to accumulate the products of density matrix elements and ERIs which make up the final operator (e.g., Coulomb or exchange). For example, the Coulomb operator can be constructed as

$$J_{\mu\nu}^{64} = P_{\lambda\sigma}^{32}(\mu\nu|\lambda\sigma)^{32} \quad (1)$$

where the superscripts indicate the number of bits of precision used for the labeled variable, and the ERIs are given as

$$(\mu\nu|\lambda\sigma) = \int \frac{\phi_{\mu}(r_1) \phi_{\nu}(r_1) \phi_{\lambda}(r_2) \phi_{\sigma}(r_2)}{|r_1 - r_2|} dr_1 dr_2 \quad (2)$$

Computing a few of the largest ERIs in full double precision has also been shown⁵ to improve accuracy compared to calculations using only single precision for all ERIs. Incremental construction of the Fock matrix¹⁷ has been noted to improve the accuracy of

Received: December 6, 2010

Published: March 23, 2011

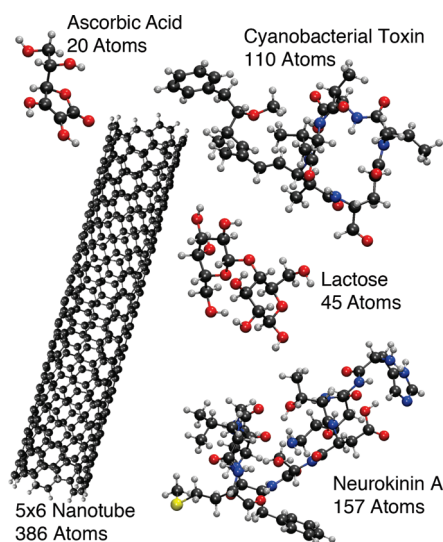


Figure 1. Molecular geometries used to benchmark the correlation between precision cutoff and the effective precision of the final energy. Optimized geometries (shown here) were used in addition to distorted nonequilibrium geometries prepared by carrying out RHF/STO-3G NVT dynamics at 2000 K (1000 K for ascorbic acid).

single precision ERI evaluation.^{3,4} Finally, it has been suggested that full single precision can be safely employed in the earliest SCF iterations.⁵ Such strategies have proven effective in improving mixed precision results for many calculations reported to date. However, no systematic study of mixed precision ERI calculation has been undertaken. In fact, the precision cutoffs must be chosen carefully for each molecular system studied to guarantee that the absolute error is within a tolerable range.

In the present paper, we introduce a systematic method of both controlling error and minimizing double precision operations in mixed precision Fock matrix construction. We begin by describing a mixed precision scheme, in which all integrals are computed on the GPU, with large ERIs calculated in full double precision and small integrals in single precision with double precision accumulation. We show that the relative error in such calculations is well behaved for an array of systems and can be controlled to provide an effective precision and performance between that of single and full double precision. In order to further decrease the number of double precision integral evaluations, we suggest a new method that we term *dynamic precision*, in which the effective precision is adjusted dynamically between SCF iterations. In this way, the minimum number of double precision integrals can be used to obtain any desired level of accuracy. Finally, we present performance and accuracy results to benchmark the dynamic precision approach. The mixed and dynamic precision schemes were implemented in TeraChem, a general purpose quantum chemistry package designed specifically for the GPU.^{11,18}

MIXED PRECISION IMPLEMENTATION

The magnitude of an ERI is commonly bounded using the Schwarz inequality.¹⁹ In direct SCF codes,¹⁷ the bound can be further reduced using elements of the density matrix as

$$|(\mu\nu|\lambda\sigma)P_{\lambda\sigma}| \leq (\mu\nu|\mu\nu)^{1/2}(\lambda\sigma|\lambda\sigma)^{1/2}|P_{\lambda\sigma}| \quad (3)$$

Because only absolute accuracy is required in chemical

Table 1. RHF/6-31G Final Energies Compared between GAMESS (Set at Default Convergence and Two-Electron Integral Thresholds), Our GPU Accelerated TeraChem Code Performing All Calculations in Double Precision (TeraChem DP), and TeraChem Using Single Precision for ERIs with Double Precision Accumulation into the Fock Matrix Elements (TeraChem SP)^a

	ascorbic acid	lactose	cyanobacterial toxin
GAMESS	−680.5828947	−1289.6666250	−2491.2058893
TeraChem DP	−680.5828947	−1289.6666250	−2491.2058890
TeraChem SP	−680.5828071	−1289.6664266	−2491.2053589
	neurokinin A		5 × 6 nanotube
GAMESS	−4089.6883770		−13790.1415171
TeraChem DP	−4089.6883762		−13790.1415176
TeraChem SP	−4089.6879824		−13790.1389987

^a Distorted nonequilibrium geometries from RHF/STO-3G NVT dynamics at 2000 K (1000 K for ascorbic acid) were used.

applications, Yasuda introduced a cutoff on the density-weighted Schwarz bound to group ERIs into two batches. Those whose bound fell below the cutoff were calculated in single precision on the GPU. Integrals whose bound was larger than the cutoff were evaluated in double precision on the CPU. As has been previously noted,¹ the use of the CPU is no longer necessary with the advent of robust double precision support on the GPU. We have implemented a mixed precision Fock matrix evaluation scheme similar to that suggested by Yasuda. Instead of using the CPU, however, we developed double precision analogues of our previously reported single precision ERI routines. Implementation details follow those in our previously described single precision code.³

In order to make the most of the GPU's memory bandwidth, the double and single precision integrals are handled in a two-pass algorithm. As previously described, our ERI algorithm operates directly on primitive Gaussian pair products, which are sorted by decreasing Schwarz bound. In the first pass, data for the largest primitive pairs are packed into double precision arrays, and any ERI whose bound is greater than the precision threshold is calculated using double precision GPU kernels. In the second pass, smaller primitive pairs are added to the data, which is reassembled into single precision arrays and processed by single precision kernels.

Because the four-index, density-weighted Schwarz bound is computed only within the GPU kernels, some duplication occurs between the sets of single and double precision primitive pairs, and each kernel must filter out individual ERIs whose bound falls outside of the relevant range. When filtering, it is essential that both the single and double precision kernels handle the bounds identically. Otherwise, the different rounding behavior exhibited by single and double precision arithmetic will cause integrals close to the bound to be skipped or double counted. In our implementation, the double precision kernels cast the bound quantities to single precision before determining whether the associated integrals and their contributions will be evaluated.

MIXED PRECISION ACCURACY

The molecules shown in Figure 1 were chosen as representative test cases to study the accuracies of several mixed precision

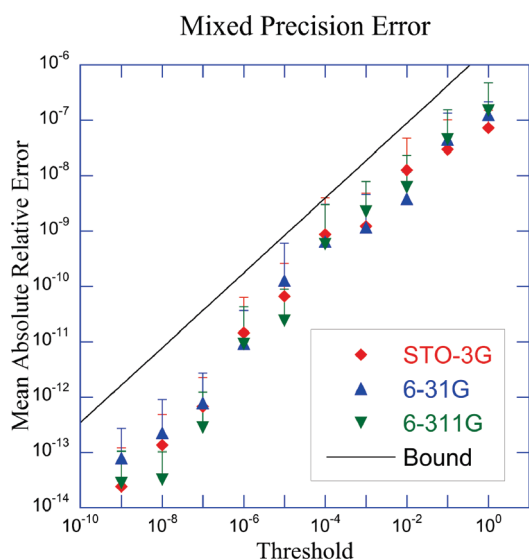


Figure 2. Average relative error in final RHF energies versus the precision threshold. Both minimized and distorted nonequilibrium geometries for the molecules in Figure 1 are included in averages. Error bars represent two standard deviations above the mean. The black line represents the empirical error bound given in eq 4.

thresholds. Geometries were prepared at both an optimized RHF/6-31G minimum and a distorted geometry obtained by performing NVT dynamics at ~ 2000 K. Coordinates are provided in Table S1 of the Supporting Information. The double precision GPU implementation was first benchmarked against GAMESS²⁰ using GAMESS default convergence and two-electron integral thresholds. The resulting final energies, shown in Table 1, indicate good agreement. Switching to single precision ERIs degrades the result by 3–5 orders of magnitude. However, it should be emphasized that for molecules containing as many as 100 atoms, single precision provides adequate results—the absolute error of the energy computed with single precision ERIs is well below 1 kcal/mol even for Neurokinin A with 157 atoms.

Next, we evaluated the mixed precision approach by varying the precision threshold between 10^{-9} au (nearly all integrals are evaluated in double precision) and 1.0 au (essentially all integrals use single precision). Negligibly small ERIs were screened according to the density weighted Schwarz upper bound of eq 3 using a conservative threshold to ensure that differences in the final energy above $\sim 10^{-7}$ au would be dominated by mixed precision errors. The average relative energy difference between full double and mixed precision SCF energies for the 10 test geometries described above is plotted as a function of the precision threshold in Figure 2. Although the absolute error increases along with the total energy of the systems, the relationship between the precision cutoff and relative error is roughly linear and is independent of system size and basis set. Thus, each threshold can be associated with an effective relative error in the energy.

When a power fit of the average errors in Figure 2 is used, it is possible to empirically preselect a precision threshold corresponding to any accuracy requirement using only the estimated total energy (to calculate relative error). A reasonably conservative bound on the precision error was obtained by shifting a power fit of the average 6-31G errors beyond two standard

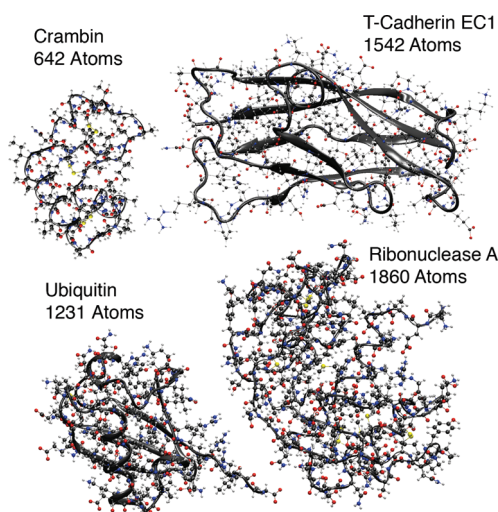


Figure 3. Additional molecules used to test the dynamic precision algorithm.

deviations above the mean. The result is given in eq 4 and plotted for reference in Figure 2.

$$\text{Err}(\text{Thre}) = 2.0 \times 10^{-6} \text{Thre}^{0.7} \quad (4)$$

By inverting eq 4, we can select an adequate precision threshold at the start of the SCF procedure and use the minimum allowable effective precision. However, this requires early iterations whose density matrices are highly approximate to use the full level of precision needed at convergence. This is especially wasteful in very large systems since the accuracy required at convergence nears that of full double precision. To further reduce the use of double precision in early iterations while still achieving the required accuracy at convergence, we introduce a dynamic precision approach, described below.

■ DYNAMIC PRECISION IMPLEMENTATION

The essence of the dynamic precision approach is to use eq 4 to select a different threshold for each iteration of the SCF procedure. Early iterations have been shown to tolerate relatively large errors in the Fock matrix without hampering convergence.^{3,21} We take the maximum element of the DIIS error vector²² from the previous iteration as a metric of this tolerance, and at each iteration, we use eq 4 to select a threshold providing precision safely below the DIIS error. This ensures that the precision error is a small contributor to the total error. By reducing the precision threshold gradually as convergence progresses, it is possible to approach full double precision results while minimizing the number of actual double precision operations.

To improve performance, our SCF code uses an iterative update approach (also known as incremental Fock matrix formation) to build up the Fock matrix over the course of the SCF procedure.¹⁷ The update approach decomposes the Fock matrix as

$$F_{i+1}(P_{i+1}) = F_i(P_i) + F(P_{i+1} - P_i) \quad (5)$$

so that only the last term needs to be calculated in each SCF iteration. Here P_i and $F_i(P_i)$ are the density and Fock matrices at the i th SCF iteration. Because changes in the density matrix become very small near convergence, the iterative Fock approach

Table 2. Comparison of Double and Dynamic Precision Final RHF/6-31G Energies (Listed in Hartree)^a

	double precision		dynamic precision		precision error	conv thres
	final energy	iter	final energy	iter		
ascorbic acid (minimum)	-680.6986413210	16	-680.6986413153	16	5.70×10^{-9}	10^{-7}
	-680.6986413213	12	-680.6986413898	12	6.85×10^{-8}	10^{-5}
ascorbic acid (1000 K)	-680.5828947151	17	-680.5828947066	17	8.50×10^{-9}	10^{-7}
	-680.5828947060	12	-680.5828947665	12	6.05×10^{-8}	10^{-5}
lactose (minimum)	-1290.0883460632	14	-1290.0883460414	14	2.18×10^{-8}	10^{-7}
	-1290.0883460086	10	-1290.0883459660	10	4.26×10^{-8}	10^{-5}
lactose (2000 K)	-1289.6666249592	15	-1289.6666249614	15	2.20×10^{-9}	10^{-7}
	-1289.6666249365	11	-1289.6666248603	11	7.62×10^{-8}	10^{-5}
cyano toxin (minimum)	-2492.3971992758	19	-2492.3971992873	19	1.15×10^{-8}	10^{-7}
	-2492.3971992730	13	-2492.3971985116	13	7.61×10^{-7}	10^{-5}
cyano toxin (2000 K)	-2491.2058890235	21	-2491.2058890017	21	2.18×10^{-8}	10^{-7}
	-2491.2058889916	13	-2491.2058886707	13	3.21×10^{-7}	10^{-5}
neurokinin A (minimum)	-4091.3672645555	19	-4091.3672645944	20	3.89×10^{-8}	10^{-7}
	-4091.3672645489	14	-4091.3672644494	14	9.95×10^{-8}	10^{-5}
neurokinin A (2000 K)	-4089.6883762179	21	-4089.6883761946	21	2.33×10^{-8}	10^{-7}
	-4089.6883760772	15	-4089.6883758130	15	2.64×10^{-7}	10^{-5}
nanotube (minimum)	-13793.7293925221	24	-13793.7293925323	23	1.02×10^{-8}	10^{-7}
	-13793.7293924922	15	-13793.7293928287	15	3.37×10^{-7}	10^{-5}
nanotube (2000 K)	-13790.1415175662	29	-13790.1415175584	27	7.80×10^{-9}	10^{-7}
	-13790.1415175332	18	-13790.1415191026	18	1.57×10^{-6}	10^{-5}
crambin	-17996.6562925538	18	-17996.6562926036	18	4.98×10^{-8}	10^{-7}
	-17996.6562925535	12	-17996.6562927894	12	2.36×10^{-7}	10^{-5}
ubiquitin	-29616.4426376594	24	-29616.4426376596	24	2.00×10^{-10}	10^{-7}
	-29616.4426376302	18	-29616.4426376655	18	3.53×10^{-8}	10^{-5}
T-cadherin EC1	-36975.6726049407	21	-36975.6726049394	21	1.30×10^{-9}	10^{-7}
	-36975.6726049265	16	-36975.6726048777	16	4.88×10^{-8}	10^{-5}
ribonuclease A	-50813.1471248227	19	-50813.1471248179	19	4.80×10^{-9}	10^{-7}
	-50813.1471247051	12	-50813.1471250247	12	3.20×10^{-7}	10^{-5}

^a Precision error is taken as the absolute difference between double and dynamic precision energies. The number of SCF iterations required to reach convergence is listed (iter) as well as the threshold used to converge the maximum element of the DIIS error matrix.

allows many additional integrals to be screened. Typically, this provides an overall speedup between $2\times$ and $3\times$ over the conventional SCF approach. However, the naïve implementation of dynamic precision described above causes the iterative Fock method to converge incorrectly, because each update of the Fock matrix does not compensate for the precision error of the previous step.

Rather than abandoning the iterative Fock algorithm altogether, we introduce the following adjustment. When the relative DIIS error drops below the error bound of the current precision threshold, the threshold is reduced to provide enough accuracy for a several orders of magnitude reduction in the DIIS error. Each time the precision is improved, the Fock matrix is recalculated from scratch. Between threshold reductions, the faster iterative Fock update scheme can be safely employed.

RESULTS

To benchmark our dynamic precision approach, we performed RHF energy calculations on the test geometries presented above, as well as some larger systems shown in Figure 3. Table 2 demonstrates the accuracy provided by our dynamic precision approach. In each calculation, the dynamic precision method is successful in reproducing the full double precision

results to within the convergence criteria. Furthermore, the number of SCF iterations required to reach convergence (also shown in Table 2) is essentially identical between dynamic and double precision. Finally, the SCF energy difference between dynamic and double precision remains fairly constant over the range of test systems, indicating that our empirical error bound is reasonably calibrated.

Of course, as with full double precision, the energy is expected to converge more rapidly than properties with first-order wave function dependence, and the final precision threshold may need to be stricter than required by eq 4. This is exactly analogous to the stricter convergence criteria that are routinely used in many quantum chemistry packages when treating first-order properties. It should be noted that in the present scheme tightening the convergence criteria will automatically reduce the final precision bound. However, a detailed analysis of precision cutoffs suitable for gradient and other property calculations is left as a topic of future research.

Table 3 summarizes the performance of our algorithm on two GPU platforms. On the older Tesla C1060 GPU, dynamic precision accelerates the SCF calculation by up to $4\times$ over full double precision. The Tesla C2050 includes a greater proportion of double precision units, and as a result the performance margin

Table 3. Runtime Comparison between Double and Dynamic Precision for RHF/6-31G Single Point Energy Calculations Converged to a Maximum DIIS error of 10^{-5} au^a

	Nvidia Tesla C1060			Nvidia Tesla C2050		
	double runtime	dynamic runtime	speedup	double runtime	dynamic runtime	speedup
ascorbic acid	7.65	2.23	3.4	4.43	2.93	1.5
lactose	36.82	9.70	3.8	15.79	8.41	1.9
cyano toxin	352.74	87.66	4.0	156.79	68.44	2.3
neurokinin A	734.61	197.91	3.7	337.68	149.76	2.3
nanotube	4693.99	1716.88	2.7	3042.92	1155.58	2.6
crambin	2754.35	1104.22	2.5	1390.49	762.09	1.8
ubiquitin	29674.48	11833.58	2.5	14997.94	7517.68	2.0
T-cadherin EC1	40936.81	17408.21	2.4	20889.98	10781.42	1.9
ribonuclease A	50092.20	21869.37	2.3			

^a Times are given in seconds and represent the wall time of the entire calculation. The hardware platform included dual Intel Xeon X5570 CPUs, 72 gigabytes of RAM, and eight GPUs. Only one GPU was used for the smaller systems (ascorbic acid through neurokinin A) to ensure that it remained saturated with parallel work. The Tesla C2050 could not treat ribonuclease A at the RHF/6-31G level due to memory constraints.

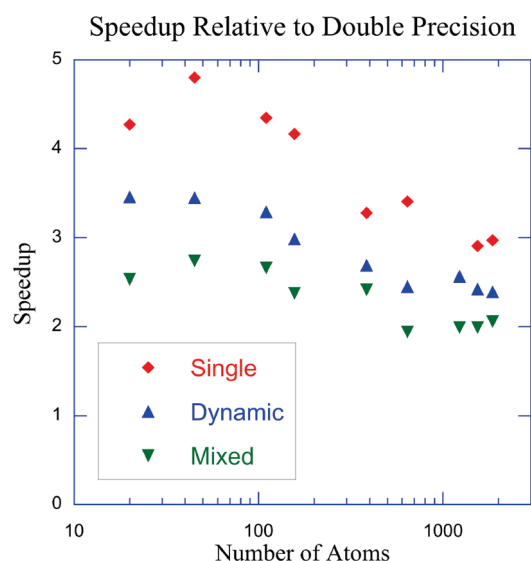


Figure 4. Speedups for RHF single point energy calculations using single, dynamic, and mixed precision relative to full double precision performance. Calculations were run on an Nvidia Tesla C1060 GPU and were converged to a DIIS error of 10^{-7} au. Mixed precision calculations used a static precision threshold chosen for each system by inverting eq 4 and solving for an absolute accuracy of 10^{-7} Hartree. Single precision failed to converge for ubiquitin.

between double and single precision arithmetic is narrowed. However, even here, dynamic precision accelerates the overall energy calculation by a full $2\times$. Figure 4 compares the wall clock speedups provided by dynamic, mixed, and single precision over full double precision calculations on the Tesla C1060 GPU. Here, *mixed precision* refers to statically fixing the precision threshold for the entire SCF procedure at the value prescribed by solving eq 4 for an error of 10^{-7} Hartrees. Dynamic precision consistently outperforms the simpler mixed precision scheme despite requiring periodic rebuilds of the Fock matrix. More importantly, dynamic precision consistently provides between 70 and 80% of the performance of single precision while providing results comparable to full double precision, and this pattern remains intact even for the largest systems.

In Figure 4, the margin between single and double precision decreases for the largest systems. This is most likely the result of the GPU's finite texture cache. In exchange kernels, density matrix elements are accessed out of order, and the texture cache is used to ameliorate noncoalesced memory loads.³ However, as the system size grows, neighboring threads begin accessing disparate parts of the density matrix and must be serviced by multiple texture loads. In the limit of large systems, the noncoalesced memory access will cause the exchange kernels to be completely memory bound, and single precision should tend to a limit $2\times$ faster than double precision owing to its smaller memory footprint.

CONCLUSION

We have demonstrated that by dynamically adjusting the ratio of integrals calculated in single and double precision on the GPU it is possible to minimize the number of double precision arithmetic operations in constructing the Fock matrix while still systematically controlling the error. Exploiting this flexibility, we have customized our Fock matrix routines for maximum performance on the GPU. Our dynamic precision implementation is able to achieve in excess of 70% of single precision's performance while maintaining accuracy comparable to full double precision. Finally, we have shown that dynamic precision is applicable to systems of unprecedented size.

The same approach can also be adapted for other hardware architectures. On traditional CPUs, for example, the use of single precision arithmetic may improve performance by reducing memory bandwidth or by enabling use of optimized SSE instructions. For extremely large systems, the required relative accuracy may well extend beyond the capacity of double precision.²³ In this limit, the approach outlined above will again prove useful in systematically improving double precision with a minimum of higher precision arithmetic operations. A more comprehensive multiprecision strategy can be easily envisioned, for example, using single, double, and quadruple precision evaluation of different ERIs, according to their magnitude. Furthermore, the same dynamical precision approach can be applied to the calculation of the Coulomb and exchange operators in density functional theory, and similar performance gains will be obtained.

■ ASSOCIATED CONTENT

S Supporting Information. Coordinates for all test molecular geometries. This information is available free of charge via the Internet at <http://pubs.acs.org/>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: todd.martinez@stanford.edu.

■ ACKNOWLEDGMENT

This work was supported by NSF (CHE-06-26354) and the Department of Defense (Office of the Director of Defense Research and Engineering) through a National Security Science and Engineering Faculty Fellowship. I.S.U. is an NVIDIA Fellow. Partial support has been provided by PetaChem, LLC through an AFOSR STTR grant administered by Spectral Sciences. T.J.M. and I.S.U. are part owners of PetaChem, LLC.

■ REFERENCES

- (1) Ufimtsev, I. S.; Martinez, T. J. *J. Chem. Theory Comput.* **2008**, *4*, 222.
- (2) Ufimtsev, I. S.; Martinez, T. J. *Comp. Sci. Eng.* **2008**, *10*, 26.
- (3) Ufimtsev, I. S.; Martinez, T. J. *J. Chem. Theory Comput.* **2009**, *5*, 1004.
- (4) Vogt, L.; Olivares-Amaya, R.; Kermes, S.; Shao, Y.; Amador-Bedolla, C.; Aspuru-Guzik, A. *J. Phys. Chem. A* **2008**, *112*, 2049.
- (5) Yasuda, K. *J. Comput. Chem.* **2008**, *29*, 334.
- (6) Olivares-Amaya, R.; Watson, M. A.; Edgar, R. G.; Vogt, L.; Shao, Y. H.; Aspuru-Guzik, A. *J. Chem. Theory Comput.* **2010**, *6*, 135.
- (7) Asadchev, A.; Allada, V.; Felder, J.; Bode, B. M.; Gordon, M. S.; Windus, T. L. *J. Chem. Theory Comput.* **2010**, *6*, 696.
- (8) Anderson, J. A.; Lorenz, C. D.; Travesset, A. *J. Comput. Phys.* **2008**, *227*, 5342.
- (9) Genovese, L.; Ospici, M.; Deutsch, T.; Mehaut, J.-F.; Neelov, A.; Goedecker, S. *J. Chem. Phys.* **2009**, *131*, 034103.
- (10) Yasuda, K. *J. Chem. Theory Comput.* **2008**, *4*, 1230.
- (11) Ufimtsev, I. S.; Martinez, T. J. *J. Chem. Theory Comput.* **2009**, *5*, 2619.
- (12) Friedrichs, M. S.; Eastman, P.; Vaidyanathan, V.; Houston, M.; Legrand, S.; Beberg, A. L.; Ensign, D. L.; Bruns, C. M.; Pande, V. S. *J. Comput. Chem.* **2009**, *30*, 864.
- (13) Harvey, M. J.; Giupponi, G.; DeFabritiis, G. *J. Chem. Theory Comput.* **2009**, *5*, 1632.
- (14) Stone, J. E.; Phillips, J. C.; Freddolino, P. L.; Hardy, D. J.; Trabuco, L. G.; Schulten, K. *J. Comput. Chem.* **2007**, *28*, 2618.
- (15) Kirk, D. B.; Hwu, W. W. *Programming Massively Parallel Processors: A Hands-On Approach*; Morgan Kaufman: Burlington, MA, 2010; p 1.
- (16) Levine, B.; Martinez, T. J. *Abstr. Pap.—Am. Chem. Soc.* **2003**, *226*, U426.
- (17) Almlof, J.; Faegri, K.; Korsell, K. *J. Comput. Chem.* **1982**, *3*, 385.
- (18) PetaChem. <http://www.petachem.com> (accessed Feb 1, 2011).
- (19) Whitten, J. L. *J. Chem. Phys.* **1973**, *58*, 4496.
- (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. J.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347.
- (21) Rudberg, E.; Rubensson, E. H.; Salek, P. *J. Chem. Theory Comput.* **2009**, *5*, 80.
- (22) Pulay, P. *J. Comput. Chem.* **1982**, *3*, 556.
- (23) Takashima, H.; Kitamura, K.; Tanabe, K.; Nagashima, U. *J. Comput. Chem.* **1999**, *20*, 443.

The Nature of the Idealized Triple Bonds Between Principal Elements and the σ Origins of Trans-Bent Geometries—A Valence Bond Study

Elina Ploshnik,[†] David Danovich,[†] Philippe C. Hiberty,^{*,‡} and Sason Shaik^{*,†}

[†]Institute of Chemistry and The Lise Meitner-Minerva Center for Computational Quantum Chemistry, The Hebrew University of Jerusalem, Jerusalem 91904, Israel

[‡]Laboratoire de Chimie Physique, UMR CNRS 8000, Université de Paris-Sud, 91405 Orsay Cédex, France

 Supporting Information

ABSTRACT: We describe herein a valence bond (VB) study of 27 triply bonded molecules of the general type $X\equiv Y$, where X and Y are main element atoms/fragments from groups 13–15 in the periodic table. The following conclusions were derived from the computational data: (a) Single π -bond and double π -bond energies for the entire set correlate with the “molecular electronegativity”, which is the sum of the X and Y electronegativities for $X\equiv Y$. The correlation with the molecular electronegativity establishes a simple rule of periodicity: π -bonding strength generally increases from left to right in a period and decreases down a column in the periodic table. (b) The σ frame invariably prefers trans bending, while π -bonding gets destabilized and opposes the trans distortion. In $HC\equiv CH$, the π -bonding destabilization overrides the propensity of the σ frame to distort, while in the higher row molecules, the σ frame wins out and establishes trans-bent molecules with $2^{1/2}$ bonds, in accord with recent experimental evidence based on solid state ^{29}Si NMR of the Sekiguchi compound. Thus, in the trans-bent molecules “less bonds pay more”. (c) All of the π bonds show significant bonding contributions from the resonance energy due to covalent–ionic mixing. This quantity is shown to correlate linearly with the corresponding “molecular electronegativity” and to reflect the mechanism required to satisfy the equilibrium condition for the bond. The π bonds for molecules possessing high molecular electronegativity are charge-shift bonds, wherein bonding is dominated by the resonance energy of the covalent and ionic forms, rather than by either form by itself.

I. INTRODUCTION

By the late 19th century, chemists recognized the diversity that the carbon–carbon multiple bonds bring into organic chemistry, and this recognition has ushered in the structural theories of Kekulé, Couper, and others, which eventually led to the development of an electronic theory of bonding by Lewis.¹ Ever since this successful chapter was opened, attempts have been made to create heavier main group analogs, for example, Si–Si double and triple bonds etc. The first attempts to synthesize compounds containing doubly bonded silicon proved, however, unsuccessful and led instead to cyclic oligomers or polymers containing only single covalent bonds.² Such failures have led to the formulation of the “Double Bond Rule”, which stated that elements having a principal quantum number greater than 2 ($n > 2$) should not be able to form π_{np-np} bonds among themselves or with other elements.³ For some years, this rule was consensual in the scientific community.

Theory has provided the rationale for the intrinsic weakness of these π_{np-np} ($n > 2$) bonds. An initial qualitative idea,⁴ that the weakness of these bonds was due to weak $np-np$ overlaps ($n > 2$), was soon refuted by Mulliken,⁵ who suggested that the rarity of these bonds is due to the larger π and σ bond strength differences in third and higher periods, leading to predominance of σ -bonded species. Later, Kutzelnigg⁶ has shown that, in fact, the propensity for π -bonding in second row elements is the exception, because these atoms lack core p orbitals, so that the 2p orbitals remain compact and result in short σ bonds. These short bond lengths, in turn, favor the π -type overlaps, leading to strong π bonds with a significant overlap population (bond orders etc).

On the other hand, np AOs of higher-row atoms develop a radial node to avoid the repulsion with the core $(n - 1)p$ AOs and are hence more diffuse than the ns AOs. This generally increases the bond lengths relative to second-row atoms, and while this lengthening does not affect the σ -type overlaps, it diminishes the π -type overlaps. Consequently in higher-row elements, σ bonds become stronger than π bonds, and multiply bonded molecules become rare. Thus, quantum chemistry provides theoretical grounds for the difficulties in making multiple bonds beyond second-row atoms, as formulated in the “Double Bond Rule”.

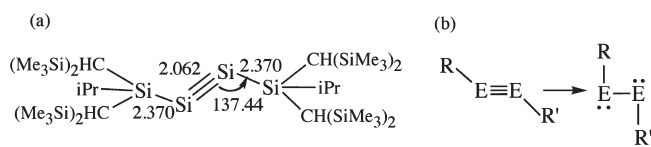
As history shows, the formulation of rules, in fact, intensifies the search for cases that break the very same rules. Indeed, the attempts to probe the limits of the “Double Bond Rule” have generated a missionary search after π -bonded elements from the third period of the periodic table and onward, and these efforts, which rose sharply in the 1980s, are still going strong and progressing to quintuple and sextuple bonding in transition metals.⁷

Already in the 1960s and 1970s, it became evident that despite the consensual acceptance of the “Rule”, transient species containing $\text{Si}=\text{C}$ and $\text{Si}=\text{Si}$ double bonds do exist. In 1981, the Rule was finally broken when the isolation of stable compounds containing $\text{P}=\text{P}$,⁸ $\text{Si}=\text{C}$,⁹ and $\text{Si}=\text{Si}$ ¹⁰ bonds was announced, which were sterically protected by bulky substituents. This strategy was used to create many more molecules containing

Received: December 23, 2010

Published: March 02, 2011

Scheme 1. (a) The Segikuchi Compound and (b) Transition between Two Bonding Motifs in REER'



double bonds between heavy main-group elements, and today almost all of the elements in periodic groups 13–16 are a part of the doubly bonded molecular family.^{11,12}

The next challenge was making triple bonds, which constitute the focus of the present paper that aims to outline the patterns of the triple bonds in the periodic table of the main elements. Thus, while the research of heavy main-group doubly bonded molecules flourished, analogous triply bonded molecules posed a more difficult challenge. Phosphaalkynes, with a P≡C triple bond, make up one group of compounds that are well-known and characterized¹³ and could be made since they do not require extremely large substituents, presumably because the P≡C bond is relatively strong. Thus, HC≡P can be kept for long periods of time at room temperature under reduced pressure,¹⁴ and t-Bu-C≡P is the first stable such molecule.¹⁵ The disulfur cation S₂I₄²⁺, isolated in S₂I₄(MF₆)₂ salts (M = As or Sb), was reported to possess a bond order between 2 and 3, according to experimental data and theoretical models.¹² But sulfur and phosphorus are exceptions, and all other heavy main group elements require very bulky substituents in order to create the triple bond.

Let us then follow with a short summary of these experimental efforts in groups 13 and 14:

The molecule OCBBCO, presumably with a B≡B bond, was spotted in reactions of B atoms with CO conducted in solid argon.¹⁵ However, no stable compound of the form [R-B≡B-R]⁻² could be prepared, and the reduction of RBX₂ compounds with bulky R ligands has led to insertion products of RB into C-H or C-C bonds.¹⁶ On the other hand, the reduction of GaCl₂C₆H₃-2,6-Trip₂ with sodium metal produced Na₂[GaC₆H₃-2,6-Trip₂]₂,¹⁷ but experimental and theoretical^{17b,18} evidence showed that the bond order is smaller than three. The geometry around the Ga≡Ga triple bond is trans-bent. No evidence for similar compounds for Al, In, or Tl was reported. Compounds with triple bonds between a group 13 atom and a group 14 or 15 element are restricted to the lightest atoms, like [R-B≡C-R']⁻, which was argued to have partial triple bond character,¹⁹ and R-B≡N,¹⁴ whose triple bond is slightly weaker than the corresponding C≡C bond.

Among the 14 group elements, stable heavier analogs of ethyne became known only recently.^{20–23} The first reports appeared in 1999 about the formation of HC≡SiCH₃, HC≡SiCl, and H₃CSi≡SiCH₃, but the evidence for these transient species was not conclusive. In 1999, Karni et al. reported²⁰ the first RC≡SiX (X = F, Cl) molecules, characterized by means of neutralization–reionization (NR) mass spectrometry. Theoretical calculations²⁰ demonstrated that these RC≡SiX species were detectable, since the barriers for isomerization to the XHC≡Si species were too high relative to the energy of the neutral vibrationally excited compound. Initial theoretical ideas for the formation of Si≡Si triple bonds focused mainly on the use of bulky R₃Si²¹ or the use of bulky aryl groups.²² In 2004, the first stable compound with a Si≡Si

bond (Scheme 1a) was isolated and fully characterized by Sekiguchi et al.,²³ who characterized it as a triple bond based on its structure, UV–vis spectrum, and calculated bond order of 2.6.²⁴

It is seen, from Scheme 1a, that the triple bond is protected by very large substituents, larger than the mesitylene groups used by West to protect the doubly bonded Si=Si compound.^{10,11} Despite the steric protection, the Si≡Si triple bond in the isolated molecule easily undergoes addition reactions with halogens. As seen from the scheme, the Si≡Si bond length is only slightly shorter than Si=Si (by 3.8%), and like the latter and its heavier analogs with Sn=Sn and Pb=Pb bonds,^{25–27} here too the Si≡Si bond is trans-bent. This is obviously in contrast with the linear structure of the acetylenes RC≡CR' molecules.

In 2001, the formation of germyne, ArGe≡CSiMe₃, with a triple bond between germanium and carbon, was reported by Bibal et al.,²⁸ as an intermediate that was trapped by alcohol solvents. This success was followed by a flurry of triply bonded heavier analogs of the Sekiguchi compound. From the structures^{29–34} of these R-E≡E-R' compounds (E = Ge, Sn, and Pb), it became apparent that they undergo a transition between triply and singly bonded motifs, as shown in Scheme 1b. Thus, Power and his co-workers²⁹ prepared R-E≡E-R' compounds of E = Si, Ge, Sn, and Pb with formal triple bonds and found that, with the exception of E = Pb, the E–E distances (and REE angles) fell in the range expected for bonds with a bond order between 2 and 3. The molecule with E = Pb, was much more bent (with an RPbPb angle ~94°) with a bond order of 1, hence resembling more a singly bonded Pb–Pb with lone pairs on the Pb centers (see Scheme 1b). Furthermore, the Sn compound showed sensitivity to the size and nature of R and R' and, with increased steric bulk around Sn, gave a highly bent (RSnSn angle ~99°) and singly bonded structure.^{29d} These findings confirmed earlier theoretical predictions^{32a–c} that the triply and singly bonded forms of REER' are not too far in energy.

Thus, while all these R-E≡E-R' compounds are trans-bent like the Sekiguchi compound shown in Scheme 1a, the bending angle is variable, as in Scheme 1b, and this makes the nature of the E–E bonding uncertain and hence vividly debated.^{30–34} A recent solid state ²⁹Si NMR experimental and theoretical study of the chemical shift tensor and chemical shift anisotropy³⁵ shows that, despite the bending, the Si≡Si bond in the Sekiguchi compound is a “genuine triple bond composed of a σ-bond and two non-degenerate π-bonds”. Accordingly, our focus on group 14 will be to understand the nature of the E≡E triple bond (E = Si and Ge) and the driving force for its bending, as in the Sekiguchi compound, while refraining from those compounds (E = Sn and Pb) which may undergo a complete transformation to the singly bonded structure (Scheme 1b).

As was mentioned already, phosphoalkynes with C≡P bonds exist.¹³ One example of arsaalkynes has been structurally characterized, Mes-C≡As (Mes = mesityl). To the best of our knowledge, no other evidence of stable nitrile analogs has been reported yet, and N₂ is the only stable homonuclear molecule with a N≡N triple bond. The corresponding heavier elements' analogues were obtained in the vapor phase. Phosphorus–nitrogen analogues of diazonium salts are also known.¹⁴ Chemists, in fact, keep turning to more candidates,^{29e} like the C≡S triply bonded molecule, recently made by Schreiner et al.³⁶

All in all, chemistry has by now a considerable family of triply bonded molecules with atoms belonging to third- and fourth-row periods, and this poses a great opportunity to generalize the nature of the triple bond in main elements, with respect to the following two features: The first is the intrinsic strength of the π bonds of the triple bond and the variation of this strength with the identity of the atoms that participate in bonding. And the second is the driving force for the distortion from linearity. Our approach is based on VB theory and is a continuation of a preliminary study,³⁷ which demonstrated that $\text{HC}\equiv\text{SiH}$ and $\text{HSi}\equiv\text{SiH}$ have 2.5 bonds, and they undergo trans bending since this deformation strongly enhances the σ bonding. The same conclusion was recently reached by Landis and Weinhold, based on natural resonance theory (NRT) calculations, leading to bond orders of 2.38–2.79 for HEEH ($\text{E} = \text{Si}, \text{Ge}, \text{and Sn}$).³⁴ As such, in the first part of the paper, we determine the *intrinsic π -bond strength* of the linear molecules and compare its dependence on the bonded atoms. Are the *intrinsic π bonds* weak? Strong? And how does this quantity vary in the periodic table. Subsequently, we intend to establish the origins of the deviation from linearity: Are these the π bonds or the σ frame which determine the tendency to distort? And how does this propensity vary with the nature of the “triply-bonded” atom?

The VB method is chosen here because it enables us to achieve these goals with a good measure of lucidity. Thus, in the VB approach, it is possible to achieve a neat separation of the σ frame and the π bonds in the molecular wave function, so that π -bond strength and their tendencies can be studied separately from the σ frame.^{37,38} As much as possible, we shall relate our results to other theoretical interpretation of these issues.

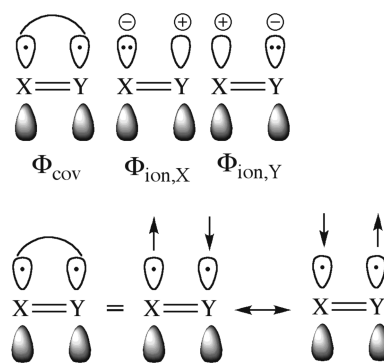
To this end, we selected a sufficiently large stock of 27 triply bonded molecules of the general form $\text{H}_n\text{X}\equiv\text{YH}_m$ ($n, m = 0, 1$; $\text{X}, \text{Y} = \text{B}^-, \text{C}, \text{N}, \text{Al}^-, \text{Si}, \text{P}, \text{Ga}^-, \text{Ge}, \text{and As}$), which will enable us to draw some generalities. We know that many of these molecules are not linear and that HSiSiH is in fact doubly bridged^{7e} and not simply bent. Still, we shall use them as linear ones, in order to establish the *intrinsic bond energies* for one of the two π bonds, $D_{2\pi}$, as well as for the total π system, $D_{2\pi}$, in the triple bonds. Having this stock of data will enable us to draw useful correlations of the intrinsic π -bonding energy with fundamental properties of the triply bonded atoms. We shall subsequently analyze the driving force for the trans-bending in the third and fourth row molecules as opposed to the linear bond in acetylene, and other second-row molecules. As shall be seen, many of these triply bonded molecule have a strong charge-shift character,³⁹ and the *total π -bonding energy correlates well with the sum of the electronegativities of the two bond constituents* rather than with the electronegativity differences.^{38a,39b,39c} Furthermore, in line with our previous findings, we shall demonstrate that invariably, within our σ - π separation, the driving force for bending in the heavier elements⁴⁰ is the strengthening of the σ bonding, which overcomes the intrinsic tendency of the π bonds to maintain a linear geometry, with the strongest possible π bonding. Thus, π bonding in heavy elements is a story of nonclassical features, when “less bonds pay more”,³⁷ and when bonds may not be classical covalent even if they look like it.

II. THEORETICAL APPROACHES AND METHODS

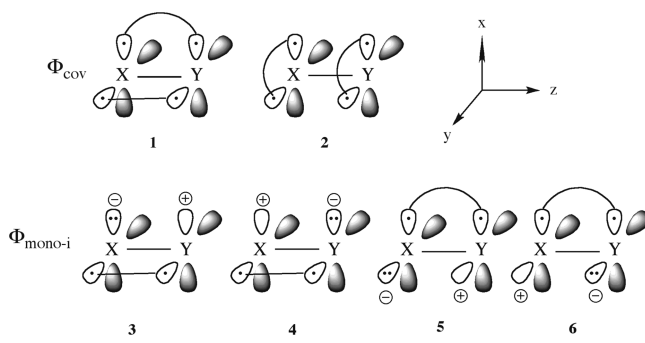
A. A Valence Bond Approach to Intrinsic Bond Energies.

The VB wave function (Ψ_{VB}), for a given number of electrons in a given set of atomic orbitals (AOs) or hybrid atomic orbitals

Scheme 2. Covalent and Ionic Structures for a π Bond and the Two Spin Arrangement Patterns Involved in Φ_{cov}



Scheme 3. Covalent and Monoionic Structures for Two π Bonds



(HAOs), is expressed as a linear combination of all of the VB structures (Φ_i) that can be generated by distributing the electrons in the AOs/HAOs so as to form a complete and linearly independent set, as expressed in eq 1:

$$\Psi_{\text{VB}} = \sum_i C_i \Phi_i \quad (1)$$

The set $\{\Phi_i\}$ is referred to as the VB-structure set.⁴¹ For the simplest case, wherein we consider a single π bond between the triply bonded X and Y, the VB-structure set consists of one covalent and two ionic structures, as shown in Scheme 2. Note that the covalent structure itself involves a combination of two spin-arrangement patterns ($\alpha\beta$ and $\beta\alpha$), and the resonance between these two structures stabilizes the covalent wave function by the spin-pairing energy, D_{cov} .

In the case of two bonds, like the two π bonds in triply bonded molecules, we have four electrons that are distributed in four AOs in all possible manners, thereby yielding a VB-structure set with 20 structures. The key ones are shown in Scheme 3, while the rest can be found in the Supporting Information (SI) document (Scheme S1). Two of these VB structures are fully covalent and are shown in Scheme 3, as 1 and 2. Compound 1 corresponds to the perfectly paired structure, while 2 pairs up the electrons on the fragments. Clearly, 2 is expected to be much less important than 1. There are also four monoionic structures, 3–6, which are generated from 1 by shifting one electron to the left or right in a single plane, either in xz or in yz , by analogy to Scheme 2 for a

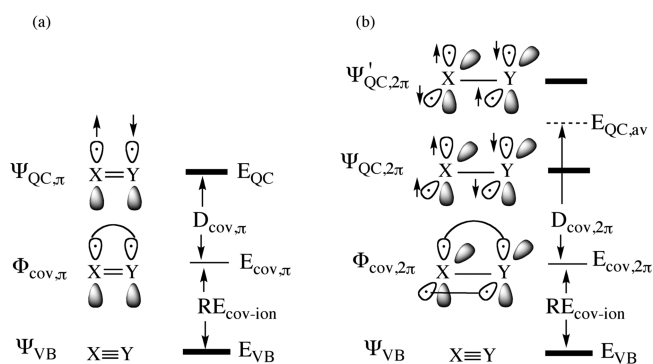


Figure 1. Definition of in situ π -bond energy and its contributing components based on the quasiclassical state (QC) reference (a) for one π bond and (b) for 2π bonds.

single bond. Other conceivable monoionics, which have the positive and negative charges in different planes, e.g., shifting electrons in **2** from p_x to p_y , are not used since they are prohibited by symmetry to mix into the wave function. In addition, there are six di-ionic structures that can be generated from the monoionics by shifting the electrons of the covalent bonds either to the left or to the right. These dionics make a very small contribution to the wave function, due to their high energy, and are therefore not shown in Scheme 3. Altogether, the VB wave function of the two π bonds is a linear combination of these 12 VB structures.

Generally speaking, whenever the covalent structure has the largest weight, which is the case in all of the molecules studied here, the bonding energy for a single bond or for a double or triple bond is given by the general eq 2:

$$\text{BDE} = \text{BDE}_{\text{cov}} + \text{RE}_{\text{cov-ion}} \quad (2)$$

Here, the total bonding energy BDE is a sum of the covalent spin-pairing energy, BDE_{cov} , and the covalent–ionic resonance energy, $\text{RE}_{\text{cov-ion}}$, due to the mixing of the ionic structures into the covalent one(s).^{39,41} Both quantities are variational within the corresponding subset of VB structures.

Determination of the Bonding Energies. For a single σ bond, one can easily determine the thermochemical bonding energy, as a difference between the molecular energy and the energy of the two dissociated radical fragments at infinity. However, for multiply bonded molecules, we cannot dissociate only the π bonds, and therefore, the thermochemical π -bonding energy cannot be determined directly. One could of course dissociate the triple bond into two fragments, e.g., 2HSi for $\text{HSi}\equiv\text{SiH}$, and determine the total bonding energy $\text{BDE}_{2\pi+\sigma}$ for the entire triple bond.³⁷ However, even then there is no clean way of separating the σ and π quantities of this $\text{BDE}_{2\pi+\sigma}$ quantity. Additionally, the thermochemical quantity involves considerable energy terms due to reorganization energies of the fragments, which in the example of HSi , involves large demotion energy from the $^4\Sigma^-$ state in the molecule to the $^2\Pi$ state in the fragment.^{29d,30,32b,32h,37,40b} Consequently, the $\text{BDE}_{2\pi+\sigma}$ quantity does not reveal the strength of those bonding interactions that are of interest here.

As we have shown amply before,^{41,42} these difficulties can be bypassed by defining a *nonbonded reference state*, in which all the π bonds are decoupled, and then determine the in situ π -bond energies as the difference between the energy of the bonded molecule relative to the nonbonded reference energy, E_{QC} . This reference nonbonded state is the quasi-classical state, $\Psi_{\text{QC},\pi}$

shown in Figure 1, when a single π bond is decoupled as in (a) and $\Psi_{\text{QC},2\pi}$ when two π bonds are decoupled as in (b).^{37,41,42}

In $\Psi_{\text{QC},n\pi}$ ($n = 1, 2$) in Figure 1, the electrons of a given π bond have only one spin arrangement pattern (only $\alpha\beta$), and therefore this structure by itself has no bonding, which arises only due to the resonance with the second spin arrangement pattern that is required to form a singlet pair (see Scheme 2). As such, the interactions across the π bonds in $\Psi_{\text{QC},n\pi}$ involve only classical electron–electron repulsion, nuclear repulsion, and electron–nuclear attraction, and since the fragments are neutral, these terms sum to approximately zero.^{6,41,42} In the case of a single π bond, as in Figure 1a, there is only one QC state, which serves as the reference. However, in the case of the two π bonds, in Figure 1b, there are two such reference states, $\Psi_{\text{QC},2\pi}$ and $\Psi'_{\text{QC},2\pi}$ which differ in the intrafragment spin relationship of the electrons in the mutually perpendicular p atomic orbitals (AOs) of the two fragments. Thus, in $\Psi_{\text{QC},2\pi}$, the two electrons on either fragment have the same spin, and hence, the X/Y fragments enjoy exchange stabilization. On the other hand, in $\Psi'_{\text{QC},2\pi}$, these electrons have opposite spins, and hence, this state is higher. To appreciate the need for these two states, we recall that upon spin-pairing of the two π bonds as in $\Phi_{\text{cov},2\pi}$ in Figure 1b, the spin in each AO will be 50% α and 50% β , and hence the proper reference state is the one having an average energy of the two QC states, $E_{\text{QC},\text{av}}$, which takes into account this averaged spin situation. Therefore, we can determine the in situ π -bond energies as a difference between the energies of the complete VB wave function Ψ_{VB} and the QC state, as follows in eq 3:

$$D_{n\pi} = D_{\text{cov},n\pi} + \text{RE}_{\text{cov-ion}}; n = 1, 2 \quad (3a)$$

$$D_{\text{cov},\pi} = E_{\text{QC}} + E_{\text{cov},\pi}, n = 1 \quad (3b)$$

$$D_{\text{cov},2\pi} = E_{\text{QC},\text{av}} + E_{\text{cov},2\pi}, n = 2 \quad (3c)$$

Here, the quantities referring to the in situ bonding energies are labeled as D , to distinguish them from thermochemical bonding energies (BDE) that are usually calculated by taking the fragments apart.

In situ Bond Energies and Driving Force for Bending. As we outlined in the Introduction, in order to make meaningful comparisons across the periodic table and to retrieve trends in π bond energies, all the molecules in the study were calculated at their linear geometries, although some of the molecules of the form $\text{H}-\text{X}\equiv\text{Y}-\text{H}$ are of trans-bent geometries in their global minimum.^{29–37} It should be emphasized that the linear geometries of those bent molecules are not random points on the PES but represent a saddle point of the second order. The establishment of intrinsic π -bond energies was then followed by determining the π -bond energies in the bent geometries and elucidating the driving forces for the distortion. Thus, the distortion energies of the σ frame were determined by calculating $\Psi_{\text{QC},2\pi}$ and $\Psi'_{\text{QC},2\pi}$ at the linear and the trans-bent geometries. Similarly, calculating D_{π} for each of the π bonds in the linear and bent geometries provides the π propensity for distortion. It is important to point out one caveat, namely, that the σ and π components in the bending plane actually mix and, hence, are not pure components. However, the variational optimization of the orbitals shows that even in the bent structures the σ frame and π bond retain these dominant characters and are in good accord with the conclusions of Kravchenko et al.³⁵ on the basis of solid

state ^{29}Si NMR of the Sekiguchi compound. Therefore, the same set of VB structures is used in the present work for the VB description of linear as well as trans-bent molecules. The validity of this choice of VB structures is further established later in the text.

An important point to note is that the QC state remains nonbonding only at distances equal to or longer than the optimal bonding distance^{43a} but becomes repulsive at shorter distances and therefore ceases to be a good reference state for gauging the in situ bonding energy. Accordingly, the method can be applied for the π components of multiple bonds but will be much less accurate for the σ component, since at a length of 1.25 Å the σ bond is highly “compressed” relative to the optimal distance for a single σ bond (~ 1.50 Å for a C–C σ bond between sp hybrids). Thus, the in situ bonding energies for the σ bonds will be estimated by a different mean, as will be detailed below.

B. Analysis of the VB Wave Functions. The weights, w_i , of the VB structures in the total wave function were determined by two methods. The first is the Coulson–Chirgwin method, which is the VB equivalent of the Mulliken population analysis, in eq 4:^{44a}

$$w_i = c_i^2 + \sum_j c_i c_j S_{ij} \quad (4)$$

The second is the inverse overlap method^{44b} in eq 5:

$$w_i \propto \frac{|c_i|^2}{(S^{-1})_{ii}}; \sum_i w_i = 1 \quad (5)$$

In both equations, the S_{ii} or S_{ij} terms are overlaps of VB configurations and c_i is the coefficient of the i th VB structure in the wave function. Since the two methods gave very similar results, the following text shows only the Coulson–Chirgwin weights, while the inverse weights are relegated to the Supporting Information (Table S1).

C. Computational Methods and Software. The linear geometries of all structures in the study were optimized by CCSD(T)/6-31G* calculations. All CCSD(T) calculations were performed with the Gaussian 03 package.⁴⁵ VB calculations were performed on the optimized linear CCSD(T)/6-31G* geometries using the XMVB-0.1 package⁴⁶ with the same basis set.

The VB methods used in this study are VBSCF and BOVB. In both methods, the orbitals are divided into two sets: an active set, made of the orbitals involved in the π central bonds, and an inactive set, involving the remaining orbitals. Core orbitals were excluded from VB calculations. The active π orbitals are AOs or HAOs, while the inactive orbitals are doubly occupied molecular orbitals. All orbitals, active and inactive, are variationally optimized to minimize the energy of the ground state. Only active orbitals are kept strictly localized in the VB sense, while the inactive ones are allowed to freely delocalize. While in VBSCF⁴⁷ all (12) VB structures share the same set of orbitals, in BOVB,⁴⁸ each structure in the wave function has its own orbital set. This allows fluctuations in the size and shape of each set of orbitals due to local charge distribution in the different VB structures, thus introducing the incremental dynamic electron correlation, during bonding, into the wave function. The BOVB method was proved to be suitable for calculations of dissociation energies.^{39b}

III. RESULTS AND DISCUSSION

The various data are displayed in tables that are arranged in a matrix form, in order to highlight the trends as both atoms move from left to right and up and down the periodic table. Each matrix

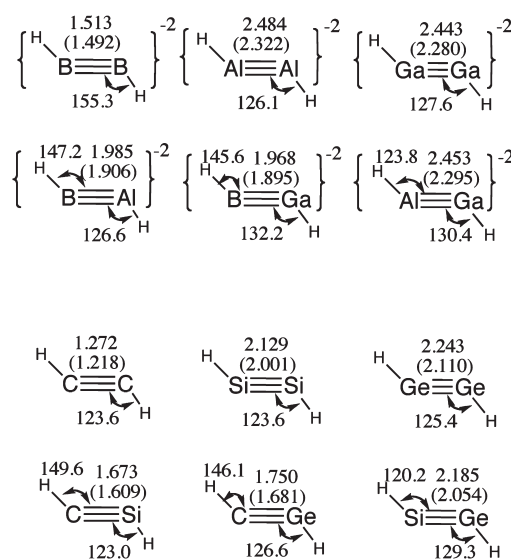


Figure 2. CCSD(T)/6-31G* optimized geometries for trans-bent molecules (distances in Å, angles in degrees). The bond lengths of the linear molecules are given in parentheses. The HCC angle in the trans-bent HC≡CH was fixed as in trans-bent HSi≡SiH while the CC distance was optimized.

is divided into submatrices, corresponding to rows of the periodic table for atoms X and Y. In the 2,2 group, X and Y are both second row atoms; in 2,3, X is second row and Y is a third row atom; and so on. In order to compare triply bonded to bent HXYH structures, the ionic configurations, X^- , of atoms B, Al, and Ga are taken, so that they can make triple bonds while being bonded to H substituents. The other atoms are neutral. Atoms C, Si, and Ge also have H substituents, while N, P, and As do not. In all tables in the following discussion, all of the molecules are labeled as $X\equiv Y$, without implicit inclusion of the hydrogen substituents or of the formal negative charges.

Geometries. The CCSD(T)/6-31G* optimized geometries of all of the molecules are shown in Figure 2 and Table 1. Figure 2 shows the fully optimized geometries for the molecules involving B, Al, Ga, C, Si, and Ge atoms, which are the only molecules in this study that can undergo trans-bending. All of these molecules, with the exception of HCCH, are more stable in the trans-bent conformation than in the linear one. To get a bent geometry for HCCH, the HCC angle was fixed to the same value as the HSiSi angle in HSiSiH, and the CC bond length was optimized with this constraint. This enables us to estimate the CC bond lengthening due to bending.

Table 1 summarizes the data for the linear molecules. The trans bending lengthens the $X\equiv Y$ distance in all cases, by widely different increments in the range 0.02–0.21 Å, where the maximal values correspond to the highest row combinations. The same trend is observed in the bending angles.

The origins of the distortion will be discussed in the end; right now, let us inspect the trends in the $X\equiv Y$ distance for the linear molecules, as summarized in Table 1.

Table 1 shows the expected grouping of the bond lengths based on the location of the X and Y atoms in the periodic table. Molecules in the group (2,2) have the shortest $X\equiv Y$ bond, while a sharp increase in the bond lengths is observed when moving to groups (2,3) and (2,4). The longest bond lengths are found in (3,3), (3,4), and (4,4), in agreement with the increased atomic size as one goes down the rows of the periodic table. Additionally,

Table 1. CCSD(T)/6-31G* X≡Y Optimized Bond Lengths (in Å) for Linear Molecules

group	X/Y	2			3			4		
		B ⁻	C	N	Al ⁻	Si	P	Ga ⁻	Ge	As
2	B ⁻	1.492								
	C		1.218							
	N		1.172	1.120						
3	Al ⁻	1.906			2.322					
	Si		1.609	1.597		2.001				
	P		1.558	1.513		1.984	1.921			
4	Ga ⁻	1.895			2.294			2.280		
	Ge		1.681	1.691		2.054	2.051		2.110	
	As		1.675	1.650		2.087	2.035		2.151	2.143

Table 2. BOVB/6-31G* Calculated Weights of Covalent and Ionic Structures for a Single π -Bond in Triply Bonded X≡Y Molecules^{a,b}

group	X/Y	2			3			4		
		B ⁻	C	N	Al ⁻	Si	P	Ga ⁻	Ge	As
2	B ⁻	0.681 <i>0.159</i>								
	C		0.638 <i>0.181</i>							
	N		0.631 <i>0.206</i>	0.625 <i>0.188</i>						
3	Al ⁻	0.649 <i>0.290</i> 0.060			0.723 <i>0.139</i>					
	Si		0.659 <i>0.168</i> 0.173	0.654 <i>0.132</i> 0.214		0.665 <i>0.167</i>				
	P		0.652 <i>0.188</i> 0.159	0.652 <i>0.159</i> 0.189		0.657 <i>0.218</i> 0.125	0.661 <i>0.169</i>			
4	Ga ⁻	0.650 <i>0.275</i> 0.075			0.717 <i>0.120</i> 0.163			0.707 <i>0.146</i>		
	Ge		0.659 <i>0.173</i> 0.169	0.659 <i>0.134</i> 0.208		0.667 <i>0.177</i> 0.155	0.663 <i>0.132</i> 0.205		0.668 <i>0.166</i>	
	As		0.650 <i>0.196</i> 0.144	0.662 <i>0.166</i> 0.172		0.660 <i>0.224</i> 0.115	0.669 <i>0.172</i> 0.159		0.667 <i>0.214</i> 0.119	0.677 <i>0.162</i>

^aThese are Coulson–Chirgwin weights (eq 4). ^bIn each cell, the topmost value in bold corresponds to the weight of the covalent structure. The other values for the ionic structures are polarized in the directions X⁻Y (italics) and X→Y (regular font).

within each group, the bond lengths decrease generally upon moving from left to right in the periodic table (e.g., descending diagonals in each group). The exception is As≡As, which possesses a longer bond compared with Ge≡Ge.

In general, these CCSD(T)/6-31G* bond lengths are in good agreement with the triple bond covalent radii, calculated by Pyykkö et al.⁴⁹ These trends are expected since the X≡Y bond lengths are determined mainly by the underlying σ bonds, which shrink along with the atomic radius from left to right of the

periodic table due to the increase in electronegativity values. However, it is notable that within each period, moving from column 13 to column 14 (B⁻→C, Al⁻→Si, Ga⁻→Ge) results in strong bond shortening, while the shortening is much less significant on transit from column 14 to column 15 (C→N, Si→P, Ge→As). One explanation is that the molecules with X,Y being B⁻, Al⁻, and Ga⁻ bear adjacent negative charges, which repel each other and tend to lengthen the triple bond. However, this electrostatic effect only concerns column 13 of the periodic

Table 3. D_{π} Bond Energies Calculated at the BOVB/6-31G* Level of Theory

group	X/Y	2			3			4		
		B ⁻	C	N	Al ⁻	Si	P	Ga ⁻	Ge	As
2	B ⁻	45.69								
	C		92.25							
	N		106.87	124.18						
3	Al ⁻	34.12			20.22					
	Si		57.32	59.68		43.93				
	P		65.48	71.32		46.59	51.29			
4	Ga ⁻	34.76			20.77			21.16		
	Ge		53.51	53.69		40.87	43.17		38.08	
	As		56.64	59.52		41.77	45.19		38.84	40.06

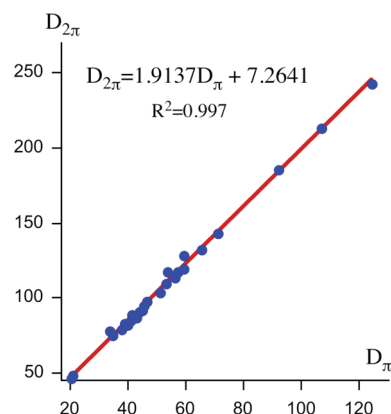
Table 4. $D_{2\pi}$ Bond Energies (in kcal/mol) Calculated at the BOVB/6-31G* Level of Theory

group	X/Y	2			3			4		
		B ⁻	C	N	Al ⁻	Si	P	Ga ⁻	Ge	As
2	B ⁻	94.68								
	C		185.15							
	N		212.65	242.81						
3	Al ⁻	77.40			44.78					
	Si		117.38	127.90		90.35				
	P		131.89	142.91		97.91	102.80			
4	Ga ⁻	74.17			46.44			47.80		
	Ge		108.82	117.23		84.48	86.80		78.99	
	As		113.45	119.33		88.26	91.16		82.32	81.14

table and does not explain the $\text{Ge}\equiv\text{Ge}\rightarrow\text{As}\equiv\text{As}$ lengthening, instead of the expected shortening. Here, another effect must be taken into account, the lone pair bond weakening effect (LPBWE) discovered by Sanderson⁵⁰ several decades ago. The LPBWE is due to overlap repulsion between the electrons of the lone pair(s) and the electrons that are coupled to a σ bond. As such, the LPBWE weakens the covalent coupling in the σ bond, resulting in bond lengthening. Since atoms of column 15 bear a lone pair while those of column 14 do not, the LPBWE explains the small bond shortenings from group (3,3) to group (4,4) and the $\text{Ge}\equiv\text{Ge}\rightarrow\text{As}\equiv\text{As}$ lengthening. We shall revisit this effect later.

Weights of the Covalent and Ionic Components of the π Bonds. Table 2 collects the BOVB/6-31G* computed weights of the covalent and ionic structures for a single π bond. It is seen that all of the bonds have dominant covalent characters. Moreover, as a general rule, bonds in the third and fourth rows are slightly more covalent than those of the second row. In each cell in the table, the covalent weight is the largest for group 13 and the smallest for group 15 (with, once again, an exception for $\text{Ge}\equiv\text{Ge}\rightarrow\text{As}\equiv\text{As}$). The behavior is similar when the two π bonds are considered, but since there are more structures, we relegate the data to the Supporting Information (Table S.2). Note that the VB structure in which both π bonds are covalent remains the predominant one, even if its weight may be inferior to 50% (as a natural consequence of the fact that the total number of VB structures is larger in the two-bond VB description than in the single-bond one).

In situ Bonding Energies for the π Bond. Table 3 displays D_{π} and Table 4 displays $D_{2\pi}$ values calculated at the BOVB/

Figure 3. A plot of BOVB/6-31G* computed D_{π} and $D_{2\pi}$ values.

6-31G* level. The VB computed $D_{2\pi}$ values are in good agreement with the trends calculated for group 14 by Frenking et al. using an energy decomposition method.^{32b} An interesting relationship between the D_{π} and $D_{2\pi}$ values is projected in Figure 3, which plots one over the other. It is seen that the slope of the correlation is close to 2, which means that the two π bonds behave approximately as two independent bonds, which in turn shows the self-consistency of the VB results.

Turning back to Tables 3 and 4, we note a few interesting patterns. As expected, the bond energy decreases down a column of the periodic table, with the unique exception of $\text{Ga}\equiv\text{Ga}$ being slightly stronger than $\text{Al}\equiv\text{Al}$. The steepest decrease occurs in

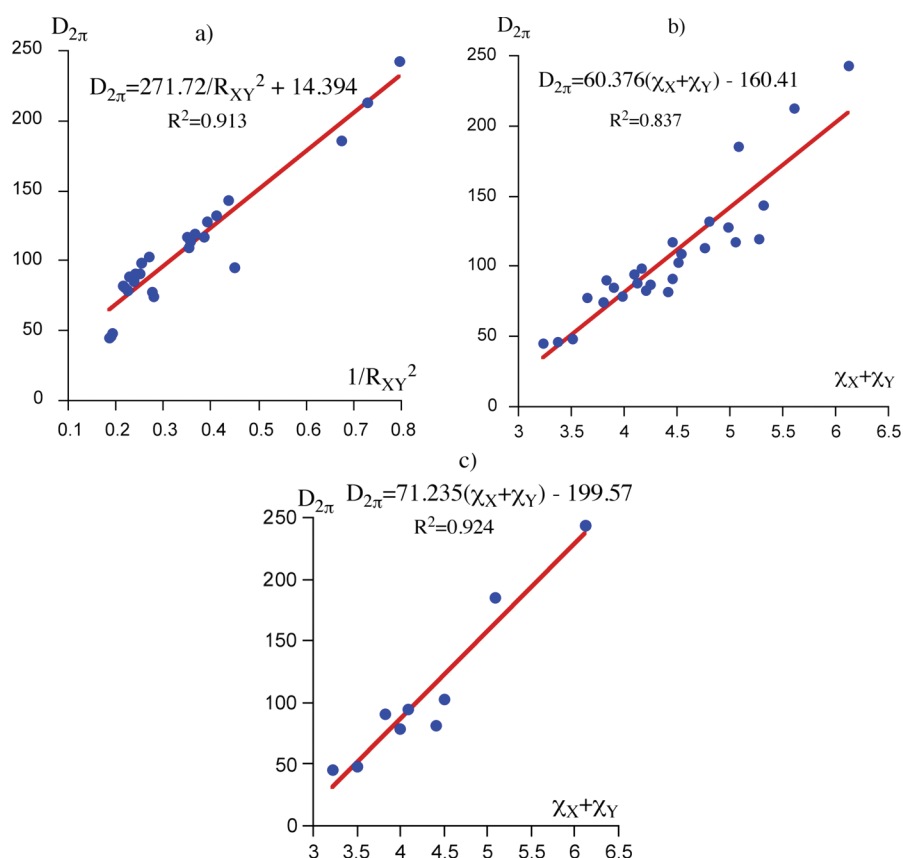


Figure 4. Plots of BOVB/6-31G* Computed $D_{2\pi}$ values against (a) the $1/R_{XY}^2$ values of the $X\equiv Y$ molecules and (b) the molecular electronegativity, $\chi_X + \chi_Y$, for the set of 27 molecules. (c) The same plot as (b) but restricted to the homonuclear bonds.

Table 5. The $D_{cov,2\pi}$ Portions (in kcal/mol) of $D_{2\pi}$ due to the Covalent VB Configuration, Φ_{cov}

group	X/Y	2			3			4		
		B ⁻	C	N	Al ⁻	Si	P	Ga ⁻	Ge	As
2	B ⁻	60.89								
	C		106.27							
	N		115.33	129.69						
3	Al ⁻	35.52			28.51					
	Si		60.67	53.71		49.40				
	P		65.59	65.53		46.40	47.71			
4	Ga ⁻	36.84			28.75			28.92		
	Ge		56.09	48.50		47.52	45.41		44.53	
	As		56.71	52.32		42.50	42.71		40.66	37.87

going from the second- to the third-row bond, and thereafter the decrease is moderate. Another clear tendency from both Tables 3 and 4 is a strengthening of the π bond(s) as one moves from left to right in a given period, hence suggesting a relationship between bond strength and electronegativity. A comparison of Tables 3 and 4 with Table 1 also shows a relationship between π -bonding strength and interatomic distance: the shorter the bond, the stronger. These trends are summarized in Figure 4: Figure 4a shows that there is a linear correlation between the $D_{2\pi}$ values and the inverse of the squared $X\equiv Y$ bond lengths. Actually, these latter quantities are related to electronegativities if one uses the definition of Allred and Rochow⁵¹ for the electronegativities of

atoms, eq 6:

$$\chi \propto Z_{\text{eff}}/R^2 \quad (6)$$

Here, R is the atomic radius and Z_{eff} is the effective nuclear charge of the atom in question.⁵¹ By this definition, the electronegativity measures the attractive force of the nuclei on the valence electrons. Extending this reasoning, one would deduce that the larger the electronegativity, the stronger the force exerted on the valence electrons of a neighboring atom and the stronger the bond between the two atoms. It is therefore logical to expect a correlation between the D_{π} or $D_{2\pi}$ values with the

Table 6. The Covalent–Ionic Resonance Energy Components of $D_{2\pi}$ in kcal/mol

group	X/Y	2			3			4		
		B [−]	C	N	Al [−]	Si	P	Ga [−]	Ge	As
2	B [−]	33.79								
	C		78.88							
	N		97.32	113.12						
3	Al [−]	41.88			16.27					
	Si		56.71	74.19		40.95				
	P		66.30	77.38		51.51	55.09			
4	Ga [−]	37.33			17.69			18.88		
	Ge		52.73	68.73		36.96	41.39		34.46	
	As		56.74	67.01		45.76	48.45		41.66	43.27

sum of electronegativities of X and Y, i.e., the “molecular electronegativity”. Such a correlation is indeed observed for the whole set of 27 molecules under study (Figure 4b), with a better correlation coefficient (0.92) when the homonuclear bonds are considered alone (Figure 4c).

Such a correlation between in situ π -bond strengths and molecular electronegativity had already been mentioned by some of us in a study of double bonds in the second and third rows.^{38a,39} In σ bonds, the same trend is observed from Li–Li to C–C but breaks down from N–N to F–F. This irregular behavior is due to the LPBWE discussed by Sanderson⁵⁰ and already mentioned above. The LPBWE weakens the σ bonds but has no direct effect on π bonds. Indeed, if the LPBWE is removed in model calculations, the so-estimated “unweakened” σ bonds show the expected linear increase from Li–Li to F–F.⁵² Thus, the linear correlation shown in Figure 4 for π bonding energies with molecular electronegativity represents a physically meaningful relationship, which in σ bonds applies only to the hypothetical “unweakened” bonds.⁵²

Components of the π -Bonding Energies. In eq 3a, above, the total bonding energy $D_{2\pi}$ for the two π bonds was expressed as a sum of two components, the $D_{\text{cov},2\pi}$ component and $\text{RE}_{\text{cov-ion}}$, the covalent–ionic resonance energy. These quantities are collected in Tables 5 and 6.

The $D_{\text{cov},2\pi}$ values in Table 5 among the different groups show nearly the same trends as the D_{π} and $D_{2\pi}$ values. The strengths of the covalent interactions generally decrease as one moves down the periodic table, especially from the second row to the third row. On the other hand, the $D_{\text{cov},2\pi}$ values for the homonuclear bonds increase in each group as one goes from left to right of the periodic table. As we argued above, this trend follows the order of molecular electronegativities, which in turn reflect the attractive force of the nuclei on the valence electrons (vide supra, eq 6). There are however two exceptions, the decrease of $D_{\text{cov},2\pi}$ from Ge≡Ge to As≡As and from Si≡Si to P≡P. Here, we recall that the As≡As bond was found to be longer than the Ge≡Ge one (Table 1), as a consequence of the lengthening effect of the LPBWE on the σ bond in As≡As. This, together with the fact that the π orbitals are necessarily more compact in As than in Ge easily explains that the covalent interaction is weaker in As≡As than in Ge≡Ge. Thus, the As≡As exception in the trends of homonuclear bonds in Table 5 is an indirect consequence of the LPBWE on the σ bonds. The same kind of explanation holds for the unexpected decrease of $D_{\text{cov},2\pi}$ from Si≡Si to P≡P.

Table 6 shows the $\text{RE}_{\text{cov-ion}}$ values, due to the mixing of the ionic structures into the covalent structures, obtained at the

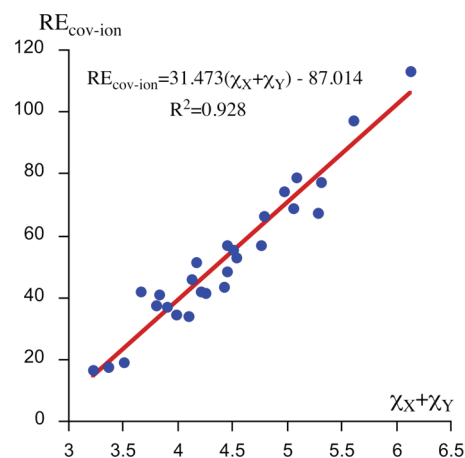


Figure 5. A plot of $\text{RE}_{\text{cov-ion}}$ vs the molecular-electronegativity value of the bond constituents.

BOVB/6-31G* level. An inspection of Table 6 shows that unlike Pauling’s scheme,⁵³ wherein it is assumed that for homonuclear bonds $\text{RE}_{\text{cov-ion}} = 0$, the present BOVB calculations show that these bonds possess very large covalent–ionic resonance energies, and the same is true for many of the other bonds in Table 6.

Figure 5 demonstrates that the molecular-electronegativity values ($\chi_X + \chi_Y$) nicely organize the $\text{RE}_{\text{cov-ion}}$ quantities for all 27 bonds. By contrast, the plot against the electronegativity difference, $\chi_X - \chi_Y$, has a very poor correlation ($r^2 = 0.17$, see Figure S1). Once again, this result is not in accord with the Pauling scheme,⁵³ wherein the $\text{RE}_{\text{cov-ion}}$ values are determined solely by the electronegativity differences.

The correlation of $\text{RE}_{\text{cov-ion}}$ with the molecular electronegativity is a fundamental relation, which is associated with the mechanism of bonding. When atoms (fragments) enter into bonding, their orbitals shrink. This shrinkage lowers the potential energy (V) of the atoms (fragments) but raises much more steeply their kinetic energies (T),⁴³ and this disrupts the virial ratio T/V which has to be -0.5 in equilibrium. As such, resonance energy is required to lower the kinetic energy and restore the equilibrium T/V ratio. When the orbitals are to begin with quite compact, as in electronegative atoms, orbital shrinkage causes too much of a kinetic energy increase, and this is further aggravated when the atoms (fragments) bear lone pairs, which raise the kinetic energy due to Pauli repulsion between them and as well as with the bonding electron. *What lowers the kinetic energy*

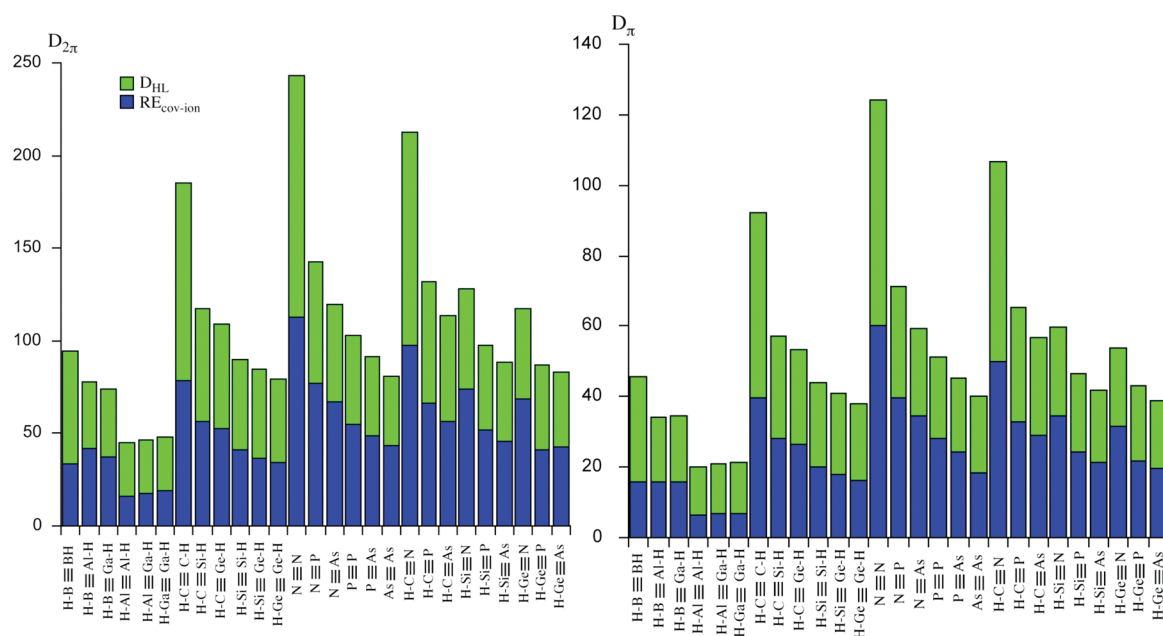


Figure 6. Diagram plots showing the D_{cov} (green) and $\text{RE}_{\text{cov-ion}}$ (blue) components of $D_{2\pi}$ (left side) and D_{π} (right side).

and restores the virial ratio to its equilibrium value is the covalent–ionic resonance energy due to the admixture of the ionic structures into the covalent structure; the more compact the atoms and the more lone pairs they have, the larger the resonance energy^{39a} that is required to restore the virial ratio. It follows that the molecular electronegativity, which is associated with either orbital compactness or the presence of lone pairs or both, favors large covalent–ionic resonance energies. This tendency, which has been demonstrated for single bonds,^{39e} has also been seen to be valid for the π components of double bonds.^{38a}

Figure 6 shows diagrammatically the total π -bond energy, and its breakdown into covalent and resonance energy contributions for both D_{π} (right side) and $D_{2\pi}$ (left side). It is apparent that, whether for one-bond or for two-bond calculations, the quantity $\text{RE}_{\text{cov-ion}}$ is always significant, and in about half of the bonds, it is either close to or more than 50% of the total bonding energy. These bonds were shown by us before to form a special family of bonds, the so-called charge-shift (CS) bonds.³⁹ Interestingly, most of the π -CS bonds in the set of 27 molecules are found when at least one of the bonded atoms involves a lone pair, e.g., in $\text{Si}\equiv\text{N}$, $\text{Ge}\equiv\text{N}$, $\text{N}\equiv\text{P}$, $\text{P}\equiv\text{P}$, $\text{As}\equiv\text{As}$, etc. This nicely illustrates the relationship between CS bonding and the presence of lone pairs in the bonded atoms, as mentioned above. It follows that the increase of π - $\text{RE}_{\text{cov-ion}}$ due to the presence of lone pair(s) compensates partially the LPBWE of the σ bonds by the same lone pairs. As an outcome, the total bonding energies of the π bonds, D_{π} or $D_{2\pi}$, always increase in a given group as the bonded atoms are taken from the left to the right of the periodic table (see also Table S3 and Figure S2 in the Supporting Information).

The Distortion Energies of $\text{HX}\equiv\text{YH}$ from Linear to Trans-Bent Forms. As can be seen in Table 1, part of the triple bonds, those containing atoms in rows higher than 2 and those involving group 13 and 14 atoms, undergo trans bending. By contrast, $\text{HC}\equiv\text{CH}$ and $\text{HC}\equiv\text{CN}$ remain linear. Following our preliminary study,³⁷ we analyzed this propensity or lack thereof as a balance between the σ and π propensities, ΔE_{σ} and ΔE_{π} . As

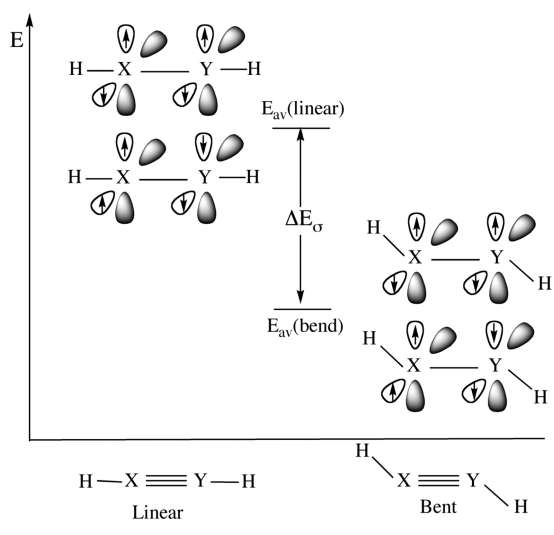
Table 7. Comparison of VB and CCSD(T) Distortion Energies (kcal/mol)^a

	CCSD(T)	VBSCF	BOVB
$(\text{HB}\equiv\text{BH})^{-2}$	2.90	0.81	2.08
$(\text{HB}\equiv\text{AlH})^{-2}$	7.88	11.28	9.38
$(\text{HB}\equiv\text{GaH})^{-2}$	5.90	9.70	8.1
$(\text{HAl}\equiv\text{AlH})^{-2}$	18.18	22.97	22.2
$(\text{HAl}\equiv\text{GaH})^{-2}$	17.57	23.58	23.11
$(\text{HGa}\equiv\text{GaH})^{-2}$	16.52	23.84	23.36
$\text{HC}\equiv\text{CH}$	−25.33	−25.85	−26.29
$\text{HC}\equiv\text{SiH}$	9.06	13.41	18.44
$\text{HC}\equiv\text{GeH}$	9.11	14.22	16.24
$\text{HSi}\equiv\text{SiH}$	23.55	30.62	28.86
$\text{HSi}\equiv\text{GeH}$	23.75	30.82	29.00
$\text{HGe}\equiv\text{GeH}$	24.09	31.77	29.79

^aThe bending energies are calculated as the energy differences between geometry-optimized bent and linear structures, as displayed in Figure 2 and Table 1, respectively.

commented upon above, such an analysis is valid if the linear and trans-bent forms of the molecule can be described by the same set of VB structures, having in each case a σ bond along the X–Y axis, and either a pair of degenerate π bonds off axis (in the linear) or an out-of plane π bond and an in-plane pseudo- π bond (in the bent). This choice of common VB structure set for the linear and bent molecule was tested by comparing the VB-calculated total distortion energies to the corresponding CCSD(T) values. The results collected in Table 7 show that VB theory predicts the distortive propensity of those molecules, which prefer the trans-bent structure. It also reveals a fair agreement between the VB and CCSD(T) sets of values, demonstrating that the VB calculations treat the linear and bent forms on equal footing and can therefore form a basis for the following discussion. Quantitative deviations between VB and CCSD(T) distortion energies may reflect that, in some cases of, e.g., the polar

Scheme 4. Definition of the σ -Propensity for Trans Bending of $\text{HX}\equiv\text{YH}$ Bonds



molecules of the heavier elements, the description of the bending process may require more structures.

The Driving Force for Trans Bending. Since the QC states are devoid of π bonds, the bending energy of the QC state will yield the σ propensity of the molecule to prefer a bent or linear structure, ΔE_{σ} . As shown in Scheme 4, the ΔE_{σ} values were calculated from the average energies of the QC states in the linear and bent geometries.

Table 8 shows the so calculated quantities for all of the molecules that can undergo trans bending, i.e., $\text{HX}\equiv\text{XY}$ with X,Y belonging to groups 13 and 14. Recall that for $\text{HC}\equiv\text{CH}$, which is linear at equilibrium, we fixed the bending angle for the bent structure as in $\text{HSi}\equiv\text{SiH}$ (123°) and reoptimized the geometry with this constraint (Figure 2).

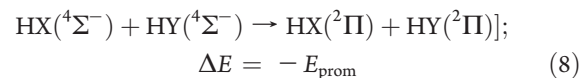
It is seen that the bonding energy of the out-of-plane π bond, $D_{\pi\text{-out}}$ is hardly affected by bending and varies within only a few kcal/mol. On the other hand, as can be seen from the variations of $D_{\pi\text{-in}}$, upon bending the in-plane pseudo- π bonds become significantly weaker in the trans-bent isomers than in the linear ones. The weakening is especially important in molecules involving atoms of group 14 (13–40 kcal/mol), and much less so in group 13 compounds (4–8 kcal/mol). As a consequence, the total bonding energy of both π bonds ($D_{\pi\text{-in}} + D_{\pi\text{-out}}$) decreases upon bending in all cases, by 6 to 46 kcal/mol.

By contrast to the π bonds, as shown by the positive ΔE_{σ} in Table 8, the σ frame is invariably stabilized by the trans bending. Further, the ΔE_{σ} values are all significant, ranging from 20 to 56 kcal/mol for group 14 and from 10 to 31 kcal/mol for group 13 compounds. Thus, all of the σ frames, including the one of acetylene, prefer the bent structure, which strengthens σ bonding. However, in all molecules displayed in Table 8, except for acetylene, this σ -bond strengthening overrides the π -bond weakening, thereby leading to trans bending. The σ propensity to distort increases for the higher row atoms, while the loss of π -bonding energy decreases in the same direction. Consequently, acetylene remains linear, despite the distortive propensity of its σ frame, while all other $\text{HX}\equiv\text{YH}$ molecules undergo trans bending. Inspection of the π -bond energy trends reveals that the in-plane bond is reduced to about 50% of its original

strength in the linear structure when X and Y belong to group 14. Hence, in agreement with previous experimental and theoretical studies,^{23,35,37} the formal number of bonds in these molecules is approximately $2^{1/2}$.

The root cause for π -bond weakening upon distortion is obviously due to the loss of overlap between the AOs involved in the in-plane pseudo- π bond as trans-bending takes place. On the other hand, the σ -bond strengthening can be due to several causes. One obvious factor is bond-length relaxation in trans-bent structures relative to linear ones (compare Figure 2 to Table 1). Thus, the bond distance between the heavy atoms in the linear structure for an optimal σ bond. Trans-bending lengthens the distance between heavy atoms, leading somewhat to σ -bond strengthening. Additionally, VB calculations using the spin-coupled method have shown that hyperconjugation also contributes to ΔE_{σ} , albeit in a minor way.³⁷ However, the most important factor for σ -bond strengthening upon bending has been shown, in a few examples (X, Y = C, Si), to be related to the increase of s orbital participation in bonding through rehybridization.^{34,37} Thus, the increase in s orbital population is $0.044 e^{-}$ for $\text{HC}\equiv\text{CH}$, $0.262 e^{-}$ for $\text{HSi}\equiv\text{CH}$, and $0.604 e^{-}$ for $\text{HSi}\equiv\text{SiH}$,³⁷ which correlates well with the ΔE_{σ} values 20, 40, and 56 kcal/mol in this series. This last effect follows the pioneering reasoning of Trinquier and Malrieu,^{40a,b} that the trans bent geometries of analogous doubly bonded molecules is due to the tendency of the molecular fragments to keep as much as possible their ground state and avoid the costly electronic promotion to the high-spin triplet states.

An Overview of σ and π Bonding in Triply Bonded $\text{HX}\equiv\text{YH}$ Molecules. The above energy separation allows us to compute the individual σ - and π -bonding contributions to the triply bonded molecules relative to the open-shell dissociated fragments. In the following approach, used before,^{32b,40,54} the $\text{HX}\equiv\text{YH}$ molecule is considered as the product of interactions between two $\text{XH} + \text{YH}$ fragments in their open shell $^4\Sigma^{-}$ states, which in most cases are the excited states of the fragments. The net dissociation energy $\text{BDE}_{\sigma+2\pi}$ is obtained after adding the demotion energy of the fragments from $^4\Sigma^{-}$ to their ground states $^2\Pi$. Thus, the dissociation process is decomposed into two formal steps, eqs 7 and 8:



where E_{prom} is the promotion energy required to excite a fragment from its $^2\Pi$ state to its $^4\Sigma$ state. Since the actual dissociation process, from the molecule to the fragments in their ground states, is the sum of reactions 7 and 8 and corresponds to the bond dissociation energy $\text{BDE}_{\sigma+2\pi}$, we can obtain D_{total} from eq 9:

$$D_{\text{total}} = \text{BDE}_{\sigma+2\pi} + E_{\text{prom}} \quad (9)$$

It is then a simple task to extract the σ contribution to bonding, D_{σ} , by subtracting the π contribution, $D_{2\pi}$, from D_{total} :

$$D_{\sigma} = D_{\text{total}} - D_{2\pi} \quad (10)$$

Table 9 displays some $\text{BDE}_{\sigma+2\pi}$ values calculated at the B3LYP/6-31G* level for a series of $\text{HX}\equiv\text{XY}$ molecules with X, Y belonging to group 14. From these values, the total bond

Table 8. σ and π Driving Forces for Bending Away from Linear $HX\equiv YH$ Molecules

	Bond Energies (kcal/mol)					
	$(H-B\equiv B-H)^{-2}$		$(H-B\equiv Al-H)^{-2}$		$(H-Ga\equiv Ga-H)^{-2}$	
	linear	bent	linear	bent	linear	bent
$D_{\pi-out}$	45.69	45.81	34.12	29.80	21.16	18.33
$D_{\pi-in}$	45.69	37.76	34.12	26.71	21.16	17.53
$(D_{\pi-in} + D_{\pi-out})$	91.38	83.57	68.24	56.51	42.32	35.86
ΔE_{π}	-7.81		-11.73		-6.46	
ΔE_{σ}	+9.89		+21.11		+29.82	
	$(H-B\equiv Ga-H)^{-2}$		$(H-Al\equiv Al-H)^{-2}$		$(H-Al\equiv Ga-H)^{-2}$	
	linear	bent	linear	bent	linear	bent
	$D_{\pi-out}$	34.76	31.31	20.22	17.19	20.77
$D_{\pi-in}$	34.76	27.11	20.22	14.12	20.77	16.35
$(D_{\pi-in} + D_{\pi-out})$	69.52	58.42	40.44	31.31	41.54	34.25
ΔE_{π}	-11.1		-9.13		-7.29	
ΔE_{σ}	19.20		31.33		30.40	
type	H-C \equiv C-H		H-C \equiv Si-H		H-Ge \equiv Ge-H	
	linear	bent	linear	bent	linear	bent
	$D_{\pi-out}$	92.25	86.44	57.32	56.03	38.08
$D_{\pi-in}$	92.25	52.01	57.32	37.19	38.08	24.94
$(D_{\pi-in} + D_{\pi-out})$	184.50	138.45	114.64	93.22	76.16	57.77
ΔE_{π}	-46.05		-21.42		-18.39	
ΔE_{σ}	19.76		39.86		48.18	
type	H-Si \equiv Si-H		H-C \equiv Ge-H		H-Si \equiv Ge-H	
	linear	bent	linear	bent	linear	bent
	$D_{\pi-out}$	43.93	37.08	53.51	51.46	40.87
$D_{\pi-in}$	43.93	23.88	53.51	36.12	40.87	24.58
$(D_{\pi-in} + D_{\pi-out})$	87.86	60.96	107.02	87.58	81.74	59.53
ΔE_{π}	-26.9		-19.44		-22.21	
ΔE_{σ}	55.76		35.68		51.21	

$^a \Delta E_{\pi} = (D_{\pi-in} + D_{\pi-out})_{\text{Bent}} - (D_{\pi-in} + D_{\pi-out})_{\text{Linear}}$; $\Delta E_{\sigma} = E_{\text{QC,av}}(\text{bent}) - E_{\text{QC,av}}(\text{linear})$. Negative values mean that the bonding is weakened upon trans bending and vice versa for positive values.

Table 9. Estimations of the Respective Bond Strengths of the σ and π Components of the Triple Bond in $HX\equiv YH$ Species, in Their Bent and Linear Conformations (All Energies in kcal/mol)

	BDE $_{\sigma+2\pi}$ ^a		E_{prom} ^b	D_{σ} ^c		$D_{2\pi}$ ^d	
	bent	linear		bent	linear	bent	linear
HC \equiv CH	201.38	231.18	40.38	103.31	87.06	138.45	184.50
HSi \equiv SiH	57.52	35.40	85.56	82.12	33.10	60.96	87.86
HGe \equiv GeH	50.77	22.26	95.68	88.68	41.78	57.77	76.16
HGe \equiv CH	99.66	91.23	68.03	80.11	52.24	87.58	107.02
HGe \equiv SiH	53.65	28.56	90.62	84.74	37.44	59.53	81.74
HSi \equiv CH	113.59	106.63	62.97	83.34	54.96	93.22	114.64

^a Bond dissociation energies calculated at the B3LYP/6-31G* level.

^b Experimental ($^2\Pi \rightarrow ^4\Sigma^-$) promotion energies, eq 8. ^c Equation 10

^d $D_{2\pi}$ values were calculated as a sum of $D_{\pi-in} + D_{\pi-out}$ energies.

strengths D_{total} are estimated through eq 9 by using experimental promotion energies D_{prom} . Finally, the D_{σ} values are estimated

through eq 10. It can be seen that in all linear conformations, the σ contribution to bonding, D_{σ} , is slightly less than one-third of the total, i.e., smaller than each of the π contributions. On the other hand, D_{σ} is clearly the strongest component of the triple bond for bent molecules.

CONCLUSION

With an aim of establishing insight into the trends in bonding and structure in a surging field of multiple bonding,^{7d,29e,30,33,35,36} we describe in this paper a valence bond (VB) study of 27 triply bonded molecules of the general type $X\equiv Y$, where X and Y are main element atoms and/or fragments from groups 13–15 in the periodic table. The molecules were studied in their linear as well as bent geometries, wherever such bending occurs (e.g., in group 14). The VB method allows the separation of π and σ bonding and thereby leads to the following conclusions:

- Both the single π -bond energy as well as the total π -bonding energy for the entire set correlate with the “molecular electronegativity”, which is the sum of the X

and Y electronegativities for $X\equiv Y$. Thus, the more electronegative the X and Y fragments, in a given triply bonded molecule, the stronger the two π bonds. The electronegativity difference is much less important. The correlation with the molecular electronegativity enables us to establish a simple rule of periodicity: Thus, following the order of the molecular electronegativity, π -bonding strength generally increases from left to right in a period and decreases down a column in the periodic table.

- (b) It was found that invariably the σ frame prefers trans bending, while π bonding is destabilized and opposes this distortion. In $\text{HC}\equiv\text{CH}$, the π -bonding destabilization overrides the propensity of the σ frame to distort, while in the higher row molecules, the σ frames win out and establish trans-bent molecules with $2^{1/2}$ bonds, in accord with recent experimental evidence based on solid state ^{29}Si NMR of the Sekiguchi compound.³⁵ Thus, as concluded before,^{34,37} the trans-bent molecules are cases where “less bonds pay more”. The distortive propensity of the σ frame originates in the increase of the s-atomic orbital population, which lowers the promotion energy ($^2\Pi \rightarrow ^4\Sigma^-$) of the fragments within the molecule.
- (c) The separation of the σ frame and π bonding allows one to determine the corresponding σ - and π -bond energies (Table 9). The trends in the σ -bond energies also follow the molecular electronegativity and increase by bending by approximately 16–49 kcal/mol for the molecules studied here.
- (d) All of the π bonds were found to have significant contributions from the resonance energy due to covalent–ionic mixing ($\text{RE}_{\text{cov-ion}}$). The $\text{RE}_{\text{cov-ion}}$ quantity for all 27 molecules correlates linearly with the corresponding molecular electronegativity, and this correlation is rooted in the bonding mechanism and the establishment of the virial ratio that typifies the equilibrium condition.^{39,43} The π -bonding energy of the more electronegative fragments has a dominant $\text{RE}_{\text{cov-ion}}$ contribution which exceeds 50% of the total bonding. Hence, all of the triple bonds have significant charge-shift character,^{39,42} and those having high molecular electronegativity are charge-shift bonds, wherein bonding is dominated by the resonance energy of the covalent and ionic forms, rather than by either form by itself.

■ ASSOCIATED CONTENT

S Supporting Information. Scheme with covalent and ionic structures for two π bonds (Scheme S1), a plot of $\text{RE}_{\text{cov-ion}}$ vs the absolute difference in molecular-electronegativity value of the bond constituents (Figure S1), BOVB/6-31G* inverse weights of covalent and ionic structures for a single π bond (Table S1), BOVB/6-31G* calculated weights (eq 4) of covalent structures for two π bonds (Table S2), a table of the $\text{RE}_{\text{cov-ion}}/D_{\text{n}\pi}$ ratio for one or two bonds Table S3, Cartesian coordinates (Table S4), and a figure of the correlation between the quantities (Figure S2). This information is available free of charge via the Internet at <http://pubs.acs.org/>.

■ AUTHOR INFORMATION

Corresponding Author

*(P.H.) Phone: +33-1-69156175. Fax: +33-1-69154447. E-mail: philippe.hiberty@u-psud.fr. (S. S.) Phone: +972-2-6585909. Fax: +972-2-6584680. E-mail: sason@yfaat.ch.huji.ac.il.

■ ACKNOWLEDGMENT

S.S. thanks the Israeli Science Foundation for a grant (ISF 53/09).

■ REFERENCES

- (1) (a) Brock, W. H. *The Norton History of Chemistry*; Norton W.W. & Co.: New York, 1992, pp 241–269; 465–483. (b) Frenking, G.; Shaik, S. J. *Comput. Chem.* **2007**, *28*, 1–3.
- (2) (a) Kipping, F. S. *Proc. R. Soc.* **1911**, *27*, 143. (b) Kipping, F. S. *J. Chem. Soc., Trans.* **1923**, *123*, 2590–2597. (c) Kipping, F. S. *J. Chem. Soc., Trans.* **1924**, *125*, 2291–2297. (d) Kipping, F. S. *Proc. R. Soc. London* **1937**, *159*, 139–148.
- (3) Jutz, P. *Angew. Chem., Int. Ed. Engl.* **1975**, *14*, 232–245.
- (4) Pitzer, K. S. *J. Am. Chem. Soc.* **1948**, *70*, 2140–2145.
- (5) Mulliken, R. S. *J. Am. Chem. Soc.* **1950**, *72*, 4493–4503.
- (6) Kutzelnigg, W. *Angew. Chem., Int. Ed. Engl.* **1984**, *23*, 272–295.
- (7) For metal–metal bonding, see for example: (a) CrCr quintuple bond in Hsu, C. W.; Yu, J. S. K.; Yen, C. H.; Lee, G. H.; Wang, Y.; Tsai, Y. C. *Angew. Chem., Int. Ed.* **2008**, *47*, 9933–9936. (b) Slighi-Dumetrescu, I.; Petrar, P.; Neme, G.; King, R. B. “Theoretical Aspects of Main Group Multiple Bonded Systems. In *Computational Bioinorganic and Inorganic Chemistry*; Solomon, E. I., Scott, R. A., King, R. B., Eds.; John-Wiley & Sons: New York, 2009; pp 563–575. (c) McGrady, J. E. Electronic Structure of Metal–Metal Bonds. In *Computational Bioinorganic and Inorganic Chemistry*; Solomon, E. I., Scott, R. A., King, R. B., Eds.; John-Wiley & Sons: New York, 2009; pp 425–432. (d) Xu, B.; Li, Q.-S.; Xie, Y.; King, R. B.; Schaefer, H. F., Jr. *J. Chem. Theory Comput.* **2010**, *6*, 735–746. (e) Frenking, G.; von Hopffgarten, M. Calculation of Bonding Properties. In *Computational Bioinorganic and Inorganic Chemistry*; Solomon, E. I., Scott, R. A., King, R. B., Eds.; John-Wiley & Sons: New York, 2009; pp 3–15.
- (8) Yoshifuji, M.; Shima, I.; Inamoto, N.; Hirotsu, K.; Higuchi, T. *J. Am. Chem. Soc.* **1981**, *103*, 4587–4589.
- (9) Brook, A. G.; Abdesaken, F.; Gutekunst, B.; Gutekunst, G.; Kallury, R. K. *Chem. Commun.* **1981**, 191–192.
- (10) West, R.; Fink, M. *J. Science* **1981**, *214*, 1343–1344.
- (11) West, R. *Angew. Chem., Int. Ed. Engl.* **1987**, *26*, 1201–1211.
- (12) Power, P. P. *Chem. Rev.* **1999**, *99*, 3463–3504.
- (13) Regitz, M. *Chem. Rev.* **1990**, *90*, 191–213.
- (14) (a) Becker, G.; Gresser, G.; Uhl, W. *Z. Naturforsch., B: Chem. Sci.* **1981**, *36*, 16–19. (b) Gier, T. E. *J. Am. Chem. Soc.* **1961**, *83*, 1769–1770.
- (15) (a) Zhou, M.; Tsumori, N.; Li, Z.; Fan, K.; Andrews, L.; Xu, Q. *J. Am. Chem. Soc.* **2002**, *124*, 12936–12937. (b) Zhou, M. F.; Jiang, L.; Xu, Q. *Chem.—Eur. J.* **2004**, *10*, 5817–5822.
- (16) (a) Mennekes, T.; Paetzold, P.; Boese, R. *Angew. Chem., Int. Ed. Engl.* **1990**, *29*, 899–900. (b) Grigsby, W. J.; Power, P. P. *J. Am. Chem. Soc.* **1996**, *118*, 7981–7988.
- (17) (a) Su, J.; Li, X.-W.; Crittendon, R. C.; Robinson, G. H. *J. Am. Chem. Soc.* **1997**, *119*, 5471–5472. (b) Xie, Y.; Grev, R. S.; Gu, J.; Schaefer, H. F.; Schleyer, P. v. R.; Su, J.; Li, X.-W.; Robinson, G. H. *J. Am. Chem. Soc.* **1998**, *120*, 3773–3780.
- (18) (a) Cotton, F. A.; Cowley, A. H.; Feng, X. *J. Am. Chem. Soc.* **1998**, *120*, 1795–1799. (b) Grunenberg, J. R.; Goldberg, N. *J. Am. Chem. Soc.* **2000**, *122*, 6045–6047. (c) Molina, J. M.; Dobado, J. A.; Heard, G. L.; Bader, R. F. W.; Sundberg, M. R. *Theor. Chem. Acc.* **2001**, *105*, 365–373.
- (19) Hunold, R.; Allwohn, J.; Baum, G.; Massa, W.; Berndt, A. *Angew. Chem., Int. Ed. Engl.* **1988**, *27*, 961–963.
- (20) Karni, M.; Apeloig, Y.; Schröder, D.; Zummack, W.; Rabezzana, R.; Schwarz, H. *Angew. Chem., Int. Ed.* **1999**, *38*, 332–335.
- (21) (a) Kobayashi, K.; Nagase, S. *Organometallics* **1997**, *16*, 2489–2491. (b) Nagase, S.; Kobayashi, K.; Takagi, N. *J. Organomet. Chem.* **2000**, *611*, 264–271.
- (22) (a) Takagi, N.; Nagase, S. *Chem. Lett.* **2001**, 966–967. (b) Kobayashi, K.; Takagi, N.; Nagase, S. *Organometallics* **2001**, *20*, 234–236.
- (23) Sekiguchi, A.; Kinjo, R.; Ichinohe, M. *Science* **2004**, *305*, 1755–1757.

- (24) Sekiguchi, A.; Ichinohe, M.; Kinjo, R. *Bull. Chem. Soc. Jpn.* **2006**, *79*, 825–832.
- (25) Fink, M. J.; Michalczyk, M. J.; Haller, K. J.; West, R.; Michl, J. *Organometallics* **1984**, *3*, 793–800.
- (26) Power, P. P. *J. Chem. Soc., Dalton Trans.* **1998**, 2939–2951.
- (27) Weidenbruch, M. *Eur. J. Inorg. Chem.* **1999**, 373–381.
- (28) Bibal, C.; Mazieres, S.; Gornitzka, H.; Couret, C. *Angew. Chem., Int. Ed.* **2001**, *40*, 952–954.
- (29) (a) (2,6-Tip₂-C₆H₃)PbPb(C₆H₃-2,6-Tip₂), Tip = 2,4,6-iPr₃C₆H₂; Pu, L. H.; Twamley, B.; Power, P. P. *J. Am. Chem. Soc.* **2000**, *122*, 3524–3525. (b) (2,6-Dipp₂-C₆H₃)SnSn(C₆H₃-2,6-Dipp₂), Dipp = 2,6-iPr₂C₆H₃; Phillips, A. D.; Wright, R. J.; Olmstead, M. M.; Power, P. P. *J. Am. Chem. Soc.* **2002**, *124*, 5930–5931. (c) (2,6-Dipp₂-C₆H₃)GeGe(C₆H₃-2,6-Dipp₂): Stender, M.; Phillips, A. D.; Wright, R. J.; Power, P. P. *Angew. Chem., Int. Ed.* **2002**, *41*, 1785–1787. (d) Fischer, R. C.; Pu, L. H.; Fettinger, J. C.; Brynda, M. A.; Power, P. P. *J. Am. Chem. Soc.* **2006**, *128*, 11366–11367. (e) Fischer, R. C.; Power, P. P. *Chem. Rev.* **2010**, *110*, 3877–3923.
- (30) For theoretical investigations of REER (E = Si–Pb) triple bonds: (a) Karni, M.; Apeloig, Y. *Chem. Isr.* **2005**, *19*, 22. (b) Karni, M.; Apeloig, Y.; Kapp, J.; Schleyer, P. v. R. In *The Chemistry of Organic Silicon Compounds*; Rappoport, Z., Apeloig, Y., Eds.; John Wiley & Sons: Chichester, U. K., 2001; Vol. 3, Chapter 1, pp 1–163. (c) Ganzer, I.; Hartmann, M.; Frenking, G. In *The Chemistry of Organic germanium, tin and lead Compounds*; Rappoport, Z., Ed.; John Wiley & Sons: Chichester, U. K., 2002; Vol. 2, Chapter 3, pp 169–282.
- (31) For experimental studies of REER (E = Ge–Pb) triple bonds: (a) Power, P. P. *Chem. Comm* **2003**, 2091–2101 and references cited therein. (b) Weidenbruch, M. *Angew. Chem., Int. Ed.* **2003**, *42*, 2222–2224. (c) Weidenbruch, M. *Angew. Chem., Int. Ed.* **2005**, *44*, 514–516. (d) Sugiyama, Y.; Sasamori, T.; Hosoi, Y.; Furukawa, Y.; Takagi, N.; Nagase, S.; Tokitoh, N. *J. Am. Chem. Soc.* **2006**, *128*, 1023–1031.
- (32) For discussions of the nature of the E–E bond in REER (M = Si, Ge, Sn), see: (a) Takagi, N.; Nagase, S. *Organometallics* **2001**, *20*, 5498–5500. (b) Lein, M.; Krapp, A.; Frenking, G. *J. Am. Chem. Soc.* **2005**, *127*, 6290–6299. (c) Jung, Y.; Brynda, M.; Power, P. P.; Head-Gordon, M. *J. Am. Chem. Soc.* **2006**, *128*, 7185–7192. (d) Grützmacher, H.; Fässler, T. F. *Chem.—Eur. J.* **2000**, *6*, 2317–2325. (e) Grunenberg, J. *Angew. Chem., Int. Ed.* **2001**, *40*, 4027–4029. (f) Malcol, N. O. J.; Gillespie, R. J.; Popelier, P. L. A. *J. Chem. Soc., Dalton Trans.* **2002**, 3333–3341. (g) Chesnut, D. B. *Heteroatom Chem.* **2002**, *13*, 53–62. (h) Sugiyama, Y.; Sasamori, T.; Hosoi, Y.; Furukawa, Y.; Takagi, N.; Nagase, S.; Tokitoh, N. *J. Am. Chem. Soc.* **2006**, *128*, 1023–1031. (j) Bridgeman, A. J.; Ireland, L. R. *Polyhedron* **2001**, *20*, 2841–2851.
- (33) (a) Pignedoli, C. A.; Curioni, A.; Andreoni, W. *ChemPhysChem* **2005**, *6*, 1795–1799. (b) Frenking, G.; Krapp, A.; Nagase, S.; Takagi, N.; Sekiguchi, A. *ChemPhysChem* **2006**, *7*, 799–800. (c) Pignedoli, C. A.; Curioni, A.; Andreoni, W. *ChemPhysChem* **2006**, *7*, 801–802.
- (34) (a) Landis, C. R.; Weinhold, F. *J. Am. Chem. Soc.* **2006**, *128*, 7335–7345. (b) Weinhold, F.; Landis, C. R. *Science* **2007**, *316*, 61–63.
- (35) Kravchenko, V.; Kinjo, R.; Sekiguchi, A.; Ichinohe, M.; West, R.; Balazs, T. S.; Schmidt, A.; Karni, M.; Apeloig, Y. *J. Am. Chem. Soc.* **2006**, *128*, 14472–14473.
- (36) Schreiner, P.; Reisenauer, H. P.; Romanski, J.; Mloston, G. *Angew. Chem., Int. Ed.* **2009**, *48*, 8133–8136.
- (37) Danovich, D.; Ogliaro, F.; Karni, M.; Apeloig, Y.; Cooper, D. L.; Shaik, S. *Angew. Chem., Int. Ed.* **2001**, *40*, 4023–4027; *Corrigenda. Angew. Chem., Int. Ed.* **2004**, *43*, 141–143.
- (38) (a) Galbraith, J. M.; Blank, E.; Shaik, S.; Hiberty, P. C. *Chem.—Eur. J.* **2000**, *6*, 2425–2434. (b) Shaik, S.; Shurki, A.; Danovich, D.; Hiberty, P. C. *Chem. Rev.* **2001**, *101*, 1501–1539.
- (39) (a) Shaik, S.; Maitre, P.; Sini, G.; Hiberty, P. C. *J. Am. Chem. Soc.* **1992**, *114*, 7861–7866. (b) Shaik, S.; Danovich, D.; Silvi, B.; Lauvergnat, D. L.; Hiberty, P. C. *Chem.—Eur. J.* **2005**, *11*, 6358–6371. (c) Shaik, S.; Danovich, D.; Wu, W.; Hiberty, P. C. *Nature Chem.* **2009**, *1*, 443–449. (d) Zhang, L. X.; Ying, F. M.; Wu, W.; Hiberty, P. C.; Shaik, S. *Chem.—Eur. J.* **2009**, *15*, 2979–2989. (e) Hiberty, P. C.; Ramozzi, R.; Song, L. C.; Wu, W.; Shaik, S. *Faraday Discuss.* **2007**, *135*, 261–272. (f) Rzepa, H. S. *Nature Chem.* **2010**, *2*, 390–393.
- (40) (a) For an early model of bonding in bent double bonds, see: Trinquier, G.; Malrieu, J. P. *J. Am. Chem. Soc.* **1987**, *109*, 5303–5315. Trinquier, G.; Malrieu, J. P. *J. Phys. Chem.* **1990**, *94*, 6184–6196. (b) Trinquier, G.; Malrieu, J.-P. *J. Am. Chem. Soc.* **1989**, *111*, 5916–5921. (c) For an early consideration of the effect of fragment states on the bond dissociation energies of double bonds, see: Carter, E. A.; Goddard, W. A. *J. Phys. Chem.* **1986**, *90*, 998–1001.
- (41) (a) Shaik, S.; Hiberty, P. C. *A Chemist's Guide to Valence Bond Theory*; Wiley-Interscience: Hoboken, NJ, 2008, pp 26–41. (b) Shaik, S.; Hiberty, P. C. *Rev. Comput. Chem.* **2004**, *20*, 1–100.
- (42) (a) Hiberty, P. C.; Danovich, D.; Shurki, A.; Shaik, S. *J. Am. Chem. Soc.* **1995**, *117*, 7760–7768. (b) Jug, K.; Hiberty, P. C.; Shaik, S. *Chem. Rev.* **2001**, *101*, 1477–1500. (c) Wu, W.; Gu, J. J.; Song, J. S.; Shaik, S.; Hiberty, P. C. *Angew. Chem., Int. Ed.* **2009**, *48*, 1407–1410. (d) Shaik, S.; Chen, Z. H.; Wu, W.; Stanger, A.; Danovich, D.; Hiberty, P. C. *ChemPhysChem* **2009**, *10*, 2658–2669.
- (43) (a) Kutzelnigg, W. The physical origin of the chemical bond. In *Theoretical Models of Chemical Bonding*; Springer: New York, 1990, part 2, pp 1–44. (b) Ruedenberg, K. *Rev. Mod. Phys.* **1962**, *34*, 326–376. (c) Feinberg, M. J.; Ruedenberg, K. *J. Chem. Phys.* **1971**, *54*, 1495–1591. (d) The role of kinetic energy in chemical binding; Wilson, C. W.; Goddard, W. A. *Theor. Chim. Acta.* **1972**, *26*, 195–210. (e) Rozendaal, A.; Baerends, E. J. *Chem. Phys.* **1985**, *95*, 57–91. (f) Ruedenberg, K.; Schmidt, M. W. *J. Comput. Chem.* **2007**, *28*, 391–410. (g) Bickelhaupt, F. M.; Baerends, E. J. *Rev. Comput. Chem.* **2000**, *15*, 1–86.
- (44) (a) Chirgwin, H. B.; Coulson, C. A. *Proc. R. Soc. London, Ser. A* **1950**, *201*, 196–209. (b) Gallup, G. A.; Norbeck, J. M. *Chem. Phys. Lett.* **1973**, *21*, 495–500.
- (45) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, Revision E.01; Gaussian, Inc.: Wallingford, CT, 2004.
- (46) Song, L.; Wu, W.; Mo, Y.; Zhang, Q. *XMVB-0.1*; Xiamen University: Xiamen, China, 2003.
- (47) van Lenthe, J. H.; Balint-Kurti, G. G. *J. Chem. Phys.* **1983**, *78*, 5699–5713.
- (48) (a) Hiberty, P. C.; Flament, J. P.; Noizet, E. *Chem. Phys. Lett.* **1992**, *189*, 259–265. (b) Hiberty, P. C.; Humbel, S.; Byrman, C. P.; Vanlenthe, J. H. *J. Chem. Phys.* **1994**, *101*, 5969–5976. (c) Hiberty, P. C.; Shaik, S. *Theor. Chem. Acc.* **2002**, *108*, 255–272.
- (49) Pyykko, P.; Riedel, S.; Patzschke, M. *Chem.—Eur. J.* **2005**, *11*, 3511–3520.
- (50) Sanderson, R. T. *Polar Covalence*; Academic Press, New York, 1983.
- (51) Allred, A. L.; Rochow, E. G. *J. Inorg. Nucl. Chem.* **1958**, *5*, 264–268.
- (52) Lauvergnat, D.; Hiberty, P. C. *THEOCHEM* **1995**, *338*, 283–291.
- (53) Pauling, L. *The Nature of the Chemical Bond*, 3rd ed.; Cornell University Press: Ithaca, 1960; pp 80–83.
- (54) Sugiyama, Y.; Sasamori, T.; Hosoi, Y.; Furukawa, Y.; Takagi, N.; Nagase, S.; Tokitoh, N. *J. Am. Chem. Soc.* **2006**, *128*, 1023–1031.

H₂-Binding by Neutral and Multiply Charged Titaniums: Hydrogen Storage Capacity of Titanium Mono- and Dications

Han Myoung Lee, Dong Young Kim, Chaeho Pak, N. Jiten Singh, and Kwang S. Kim*

Center for Superfunctional Materials, Department of Chemistry, Pohang University of Science and Technology, San 31, Hyojadong, Namgu, 790-784 Pohang, South Korea

ABSTRACT: Given that transition metal–hydrogen systems have been studied as a predecessor for hydrogen storage materials, we have investigated the neutral and multiply charged titanium–H₂ systems (Ti–H₂, Ti⁺–H₂, Ti²⁺–H₂, Ti³⁺–H₂, and Ti⁴⁺–H₂) using density functional theory (DFT) and high-level ab initio calculations, including coupled cluster theory with single, double, and perturbatively triple excitations [CCSD(T)]. These systems show different types of hydrogenation depending on their charged state. The neutral Ti–H₂ system shows dihydride structure with covalent interaction where the Ti–H distance is 1.76 Å, while H₂ is dissociated into two neighboring hydride ions by withdrawing electrons from Ti. The charged Ti⁺–H₂, Ti²⁺–H₂, and Ti³⁺–H₂ systems show dihydrogen structures with noncovalent interaction, where the Ti⁺–H, Ti²⁺–H, and Ti³⁺–H distances are 2.00, 2.14, and 2.12 Å, respectively. The main binding energies in these systems arise from the hydrogen polarizability driven interaction by the positive charge of Tiⁿ⁺ (n = 1–3). Among Tiⁿ⁺–H₂ (n = 1–3) the Ti⁺–H₂ has the shortest distance against our common expectation, while Ti²⁺–H₂ has the longest distance. The Ti⁺–H₂ distance is the shortest because of the d–σ* molecular orbital (MO) interaction which is not present in Ti²⁺–H₂ and Ti³⁺–H₂. The Ti⁴⁺ ion does not bind H₂. In this regard, we have investigated the maximal hydrogen binding capacity by Ti complexes. The coordination of titanium mono- and dications complexed with dihydrogen (H₂) [Ti⁺(H₂)_n and Ti²⁺(H₂)_m] is studied along with their structures, binding energies, electronic properties, and spectra. The titanium monocations of the quartet ground state have up to the hexacoordination, while titanium dications of the triplet ground state have up to the octacoordination at very low temperatures. At room temperature, the monocations favor penta- to hexacoordination, while the dications favor hexacoordination. This information would be useful for the design of hydrogen storage devices of Ti complexes, such as Ti-decorated/dispersed polymer–graphene hybrid materials.

I. INTRODUCTION

Development of the safe, reliable, compact, and cost-effective hydrogen storage medium is a technically challenging issue facing the “hydrogen economy”.¹ Two major techniques for hydrogen storage is to store hydrogen in the atomic form of hydrides and in the molecular form in sorbents. Some of metal hydrides, complex hydrides,² simple hydrogen-containing chemicals,³ and mixtures of these⁴ have shown promising results as hydrogen capacious materials at ambient conditions. They not only meet the ultimate goal of the system capacity with 7.5 wt % H₂ set by the U.S. Department of Energy (DOE)⁵ but also release hydrogen at ~363 K, which is the operating temperature of proton exchange membrane (PEM) fuel cells. However, no materials explored to date exhibit practical utility. On the other hand, hydrogen sorbents, such as carbon-based nanostructures and metal/covalent organic frameworks, show fast adsorption and desorption kinetics.⁶ However, because they are conventionally based on physisorption by the weak van der Waals interactions, they store hydrogen at high pressures and very low temperatures. While many researchers have been trying to strengthen the interaction between the sorbents and hydrogen, the use of the Kubas-typed orbital interaction (metal–σ interaction)⁷ would be an ideal strategy to make new generation of hydrogen sorbents.

Ti-decorated organometallic systems are suggested as good hydrogen storage nanomaterials through theoretical study.⁸ Transition-metal–π complexes (especially titanium complexes)⁹ and titanium metal–organic frameworks are also suggested as a

H₂ storage material.¹⁰ Porphyrins incorporated with the first-row transition metals were reported for their binding energies with a hydrogen molecule,¹¹ where the binding energy by Ti was the largest. The adsorption of H₂ on a series of 3d transition-metal (TM)-doped organosilica complexes was investigated by using theoretical methods.¹² The modified benzene–silica model with Sc, Ti, V, Cr, and Mn transition-metal atoms could adsorb H₂ into the dihydride or dihydrogen configuration forms. The structural, energetic, and electronic properties of small hydrogenated titanium clusters were studied.¹³ Their electronic structures were dealt with as a function of the number of adsorbed hydrogen atoms, which is an important issue in nanocatalysis and hydrogen storage. Given that the information of the structures and coordination of molecular clusters has been utilized to design novel functional materials,¹⁴ we have been interested in investigating the structure and coordination of neutral and multiply charged hydrogenated transition metals for possible design of organometallic compounds useful for hydrogen storage materials.

Transition metals have long been investigated for their catalytic properties in the hydrogenation reactions. Thus, their dihydrogen–metal complexes, where H₂ molecules were bound as intact molecular ligands, were investigated. The cation–(H₂)_n clusters with various first-row transition metal ions, such as Sc,¹⁵ Ti,¹⁶ V,¹⁷ Cr,¹⁸ Mn,¹⁹ Fe,²⁰ Co,²¹ Cu,²² and Zn,¹⁹ were reported.

Received: December 23, 2010

Published: March 16, 2011

Table 1. Binding Energies, Selected Geometrical Parameters, NBO Charges of Ti, and Selected Frequencies of $\text{Ti}^{0/+2+/3+}-\text{H}_2$ Complexes^a

method	$-\Delta E$	$-\Delta E_0$	$-\Delta H_r$	$-\Delta G_r$	$d_{\text{H-H}}$	$d_{\text{Ti-H}}$	$q(\text{Ti})$	ν_{H_2}	$\nu_{\text{Ti-H}_2}^{\text{asym}}$	$\nu_{\text{Ti-H}_2}^{\text{sym}}$
Ti-H ₂										
B3LYP/W*QZ	16.47	17.40	18.47	13.53	3.08	1.76	—	(573) ^b	1593	1582
MP2/W*QZ	10.61	11.80	12.83	7.89	3.18	1.75	1.222	(458) ^b	1568	1559
CCSD(T)/W*QZ	11.82	13.03	14.04	9.15	3.19	1.76	—	(436) ^b	1572	1543
Ti ⁺ -H ₂										
B3LYP/W*QZ	11.60	9.90	11.00	5.78	0.77	2.03	—	3839	1036	712
MP2/W*QZ	11.75	9.89	10.99	5.76	0.76	2.00	1.031	3969	1013	683
CCSD(T)/W*QZ	9.53	7.70	8.80	3.55	0.77	2.00	—	3904	1057	718
Ti ²⁺ -H ₂										
B3LYP/W*QZ	20.50	19.03	20.09	15.00	0.78	2.17	—	3958	784	686
MP2/W*QZ	17.88	16.31	17.37	12.28	0.76	2.16	1.978	4024	770	675
CCSD(T)/W*QZ	19.20	17.57	18.63	13.51	0.77	2.14	—	4037	783	720
Ti ³⁺ -H ₂										
B3LYP/W*QZ	—	—	—	—	—	—	—	—	—	—
MP2/W*QZ	53.26	52.79	53.79	48.77	0.83	2.05	2.827	3372	947	455
CCSD(T)/W*QZ	56.94	57.37	58.51	53.35	0.84	2.12	—	3226	903	159

^a ΔE and ΔE_0 are the binding energies (kcal/mol) without/with zero-point energy correction; ΔH_r and ΔG_r are the enthalpy and free-energy changes (kcal/mol) at room temperature and 1 atm; $d_{\text{H-H}}$ is H-H distance (Å); and $d_{\text{Ti-H}}$ is the distance (Å) between a Ti ion and a H atom of H₂. The H-H distance of pure H₂ molecule is 0.742, 0.736, and 0.742 Å at the B3LYP, MP2, and CCSD(T) levels, where the aug-cc-pVQZ set was used for hydrogen and the [9s7p5d3fg] contracted basis set (W*) was used for Ti; and $q(\text{Ti})$ is the natural bond orbital charge (au) of the Ti ion at the MP2 level. The NBO charges are given at the MP2 level; ν_{H_2} is the scaled stretching frequency (cm⁻¹) of the H₂ molecule; and $\nu_{\text{Ti-H}_2}^{\text{asym}}$ and $\nu_{\text{Ti-H}_2}^{\text{sym}}$ are scaled asymmetric and symmetric Ti-H₂ stretching frequencies (cm⁻¹). The experimental H-H stretching frequency of the H₂ molecule is 4401 cm⁻¹ (ref 28). ^b Bending frequencies of Ti-H₂.

Their objectives were to obtain the structures and binding energies of metal ion-(H₂)_n complexes and to understand the σ -bond activation processes for the H-H bond and for the C-H bonds in hydrocarbons. Since the binding of Ti⁺ with H₂ was predicted to be the largest among first transition metals,¹⁶ we here investigate dihydrogen-titanium complexes. The charged state of titanium atom plays an important role in complexation.²³ We report the structures, binding energies, and electronic properties of titanium mono- and dications binding dihydrogen (H₂) molecules [Ti⁺(H₂)_{n=1-7} and Ti²⁺(H₂)_{m=1-9}] using density functional theory (DFT) and high-level ab initio theory.

II. CALCULATION METHODS

For all Möller-Plesset second-order perturbation (MP2) and coupled cluster theory with single, double, and perturbatively triple excitations [CCSD(T)] calculations presented here, all electrons were correlated, the unrestricted open shell approach was employed, and the basis set superposition error (BSSE) correction was made. The calculations were carried out by using the Gaussian 03 suite of programs.²⁴ Molecular structures and molecular orbital contour maps were drawn using the Posmol package.²⁵

For Tiⁿ⁺-H₂ ($n = 0-4$), we have used DFT with Becke's three-parameters for exchange and Lee-Yang-Parr correlation functional (B3LYP) in the investigation of neutral and cationic hydrogen-titanium clusters. The aug-cc-pVQZ set is employed for hydrogen. By extending the Ti basis set of the Wachters+f set [8s6p4df] (close to the ccpVTZ and aug-cc-pVDZ levels, which will be denoted as W),²⁶ we have made the optimized [9s7p5d3fg]

contracted functions (similar to ccpVQZ and aug-cc-pVTZ levels, which will be denoted as W*). All the "d" and "f" orbitals are the spherical harmonic basis functions (5d and 7f). The combined basis set of W and aug-cc-pVTZ will be simply denoted as WTZ for simplicity, and that of W* and aug-cc-pVQZ will be denoted as W*QZ. Since B3LYP results are not reliable in the present system, the important structures were calculated at the MP2 level using the same basis sets. To obtain more accurate results, the CCSD(T)/W*QZ calculations were employed. At the CCSD(T)/W* level the initial $\langle S^2 \rangle$ values of Ti^{0/+2+/3+} are 2.010, 3.753, 2.003, and 0.752, and the T1 diagnostic values of Ti^{0/+2+/3+/4+} are 0.0145, 0.0049, 0.0023, 0.0015, and 0.0008, respectively. The spin multiplicities of Ti^{0/+2+/3+/4+} are triplet, quartet, triplet, doublet, and singlet, respectively. The CCSD(T)/W* ionization energies for Ti^{+2+/3+/4+} are 151.1, 309.5, 632.4, and 999.3 kcal/mol, in agreement with the experimental values (157.3, 313.1, 634.1, and 997.8 kcal/mol).²⁷ At each calculation level the structures were fully optimized along with vibrational frequencies. The frequencies were employed to obtain the zero-point energies (ZPE) and the enthalpy/free-energy changes ($\Delta H_r/\Delta G_r$) at room temperature and 1 atm. To investigate their stabilities, the binding energies of ΔE_c and ΔE_0 between Ti ions and H₂ were calculated without and with ZPE correction, respectively. The frequencies were scaled by scale factors (0.997 at the B3LYP/W*QZ, 0.975 at the MP2/W*QZ, and 1.000 at the CCSD(T)/W*QZ level) to match the experimental frequency (4401 cm⁻¹)²⁸ of the H₂ molecule. Though the neutral Ti-H₂ complex forms a covalent bond between a neutral Ti atom and two hydrogen atoms, multicationic charged complexes show noncovalent interactions between a Ti ion and a

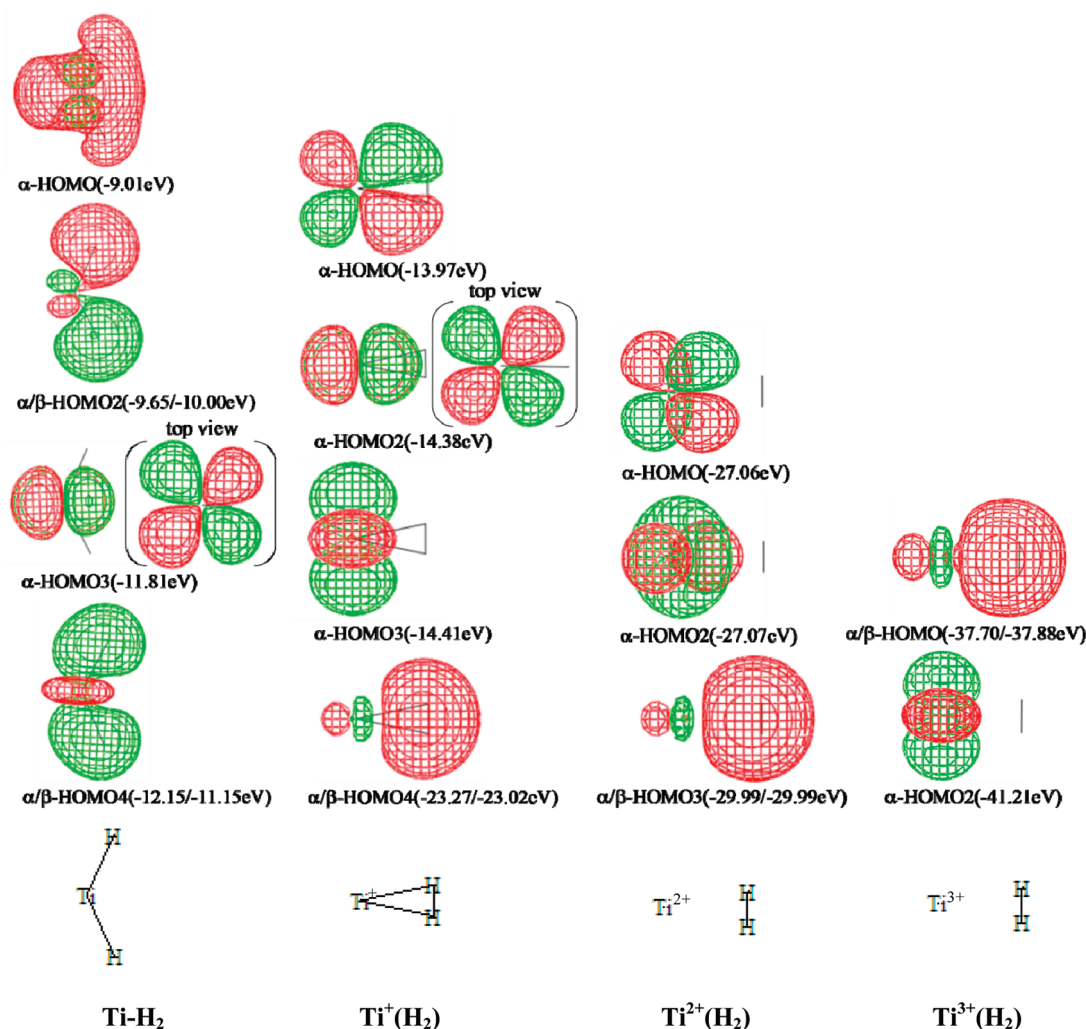


Figure 1. HOMOs of the $\text{Ti}-\text{H}_2$, Ti^+-H_2 , $\text{Ti}^{2+}-\text{H}_2$, and $\text{Ti}^{3+}-\text{H}_2$ complexes (at the CCSD(T)/WTZ level). The orbital energies are in parentheses.

hydrogen molecule. The natural bond orbital (NBO) charges were obtained at the MP2/W*QZ level of theory.

We investigated various possible structures of $\text{Ti}^+(\text{H}_2)_{n=1-7}$ and $\text{Ti}^{2+}(\text{H}_2)_{m=1-9}$. The low-energy structures were initially investigated by B3LYP/WTZ and further calculated by the MP2/WTZ. The structures were fully optimized along with their vibrational frequency analysis. The CCSD(T)/WTZ calculations were performed on the MP2/WTZ optimized geometries. At the CCSD(T)/W level the initial $\langle S^2 \rangle$ values of Ti^+ and Ti^{2+} ions were 3.75 and 2.00, and their T1 diagnostic values were 0.0060 and 0.0040, respectively. At the CCSD(T)/W level the energy difference between Ti^+ and Ti^{2+} ions was estimated to be 306.9 kcal/mol, in agreement with the experimental value (313.1 kcal/mol).²⁷ The CCSD(T)/WTZ thermodynamic quantities (ΔE_0 at 0 K; ΔH_r and ΔG_r at room temperature and 1 atm) were estimated using the CCSD(T)/WTZ single point interaction energies (ΔE_e) and the MP2/WTZ ZPE and thermal energies. The NBO charges (q^{NBO}) were estimated through NBO population analysis at the MP2/WTZ level of theory. Their B3LYP/WTZ and MP2/WTZ frequencies were scaled by scale factors (0.966 and 0.975) to match the experimental frequency (4401 cm^{-1})²⁸ of the H_2 molecule.

III. RESULTS AND DISCUSSION

The binding energies, selected geometrical parameters, NBO charges of Ti [$q(\text{Ti})$], and selected frequencies of $\text{Ti}^{0/1+/2+/3+}-\text{H}_2$ complexes, which were calculated at the levels of B3LYP, MP2, and CCSD(T) using the W*QZ basis set, are listed in Table 1. The $q(\text{Ti})$ values of the $\text{Ti}^{0/1+/2+/3+}-\text{H}_2$ complexes are 1.22/1.03/1.98/2.83 au. Figure 1 shows the frontier highest occupied molecular orbitals (HOMO) of $\text{Ti}-\text{H}_2$, Ti^+-H_2 , $\text{Ti}^{2+}-\text{H}_2$, and $\text{Ti}^{3+}-\text{H}_2$ complexes at the CCSD(T) level with the WTZ basis set, for visual aid, because those with W*QZ are hard to be legible due to the diffuse nature of the orbitals.

The neutral $\text{Ti}-\text{H}_2$ complex shows the covalent-bonded Ti-dihydride structure. The spin multiplicity for the ground state of the $\text{Ti}-\text{H}_2$ complex is triplet. It has a bent shape like the water molecule (H_2O). The binding energy without/with ZPE correction ($-\Delta E_e/\Delta E_0$) is 16.47/17.40, 10.61/11.80, and 11.82/13.03 kcal/mol at the B3LYP/W*QZ, MP2/W*QZ, and CCSD(T)/W*QZ levels. As for the HOMO, the α -HOMO is composed of the hybridized orbital of 4s and 3d orbitals of Ti and the σ orbital of H_2 . A doubly occupied d orbital of Ti and the σ^* orbital of H_2 form the second α/β -HOMO (α/β -HOMO2), which leads to the strong backdonation to the $\text{H}_2\sigma^*$ orbital,

inducing the bent structure. Thus, the NBO charge (q) of H atoms is negative and that of Ti atom is positive ($q_{\text{Ti}} = 1.22$ au, $q_{\text{H}} = -0.61$ au at the MP2/W*QZ level). The third α -HOMO (α -HOMO3) is a singly occupied nonbonding d orbital. The fourth α/β -HOMO (α/β -HOMO4) is a bonding orbital between an unoccupied d_{z^2} orbital and a doubly occupied $\text{H}_2\sigma$ orbital.

At the B3LYP/W*QZ level the complex has a relatively weak orbital interaction between an occupied d orbital and the unoccupied $\text{H}_2\sigma^*$ orbital, which leads to the very strong $d-\sigma^*$ backdonation as compared with the cases at the MP2/W*QZ and CCSD(T)/W*QZ levels. This backdonation is due to the third α/β -HOMO (α/β -HOMO3) at the B3LYP/W*QZ level, the second α/β -HOMO (α/β -HOMO2) at the CCSD(T)/W*QZ level, and the α/β -HOMO at the MP2/W*QZ level. Therefore, the B3LYP/W*QZ binding energy is larger than the MP2/W*QZ value. This effect results in slightly different bond angles: 122.3° at the B3LYP/W*QZ level, 129.7° at the CCSD(T)/W*QZ level, and 131.0° at the MP2/W*QZ level.

The spin multiplicities of Ti^+-H_2 , $\text{Ti}^{2+}-\text{H}_2$, and $\text{Ti}^{3+}-\text{H}_2$ are quartet, triplet, and doublet, respectively. The Ti^+-H_2 complex has the $d-\sigma^*$ backdonating bonding orbital (α -HOMO), which does not appear in the occupied regions of the $\text{Ti}^{2+}-\text{H}_2$ and $\text{Ti}^{3+}-\text{H}_2$ complexes as shown in Figure 1. Therefore, the Ti–H distance of the Ti^+-H_2 complex is shorter than those of $\text{Ti}^{2+}-\text{H}_2$ and $\text{Ti}^{3+}-\text{H}_2$ complexes as shown in Table 1. The NBO charges of H atoms (q_{H}) are -0.015 au. In the case of the Ti^+-H_2 complex the α -HOMO is a bonding orbital between a singly occupied d orbital and the unoccupied $\text{H}_2\sigma^*$ orbital. This orbital has the $d-\sigma^*$ backdonation interaction. The second and third α -HOMOs (α -HOMO2 and α -HOMO3) are singly occupied nonbonding d orbitals. The fourth α -HOMO and β -HOMO (α/β -HOMO4) have the weak bonding character between an unoccupied d orbital and the doubly occupied σ orbital of H_2 . This orbital interaction also appears as the third α/β -HOMO (α/β -HOMO3) in $\text{Ti}^{2+}-\text{H}_2$ and as the α/β -HOMO in the $\text{Ti}^{3+}-\text{H}_2$. The α -HOMO and α -HOMO2 of $\text{Ti}^{2+}-\text{H}_2$ are nonbonding orbitals.

The Ti– H_2 complex has two Ti–H stretching modes (asymmetric and symmetric modes) and one bending mode. As shown in Table 1, the scaled frequency values of asymmetric/symmetric stretching modes at the B3LYP/W*QZ, MP2/W*QZ, and CCSD(T)/W*QZ levels are 1593/1582, 1568/1559, and 1572/1543 cm^{-1} , respectively; those of the bending mode are 573, 458, and 436 cm^{-1} , respectively; the Ti–H distances are 1.759, 1.748, and 1.762 Å, respectively.

The $\text{Ti}^{+2+/3+}-\text{H}_2$ complex has the dihydrogen bond structure by the noncovalent interaction between a Ti ion and an H_2 molecule. In the $\text{Ti}^{4+}-\text{H}_2$ system the electron of H_2 molecule transfers to the Ti^{4+} ion, and the q_{H} is greater than 0.5 au when the H is ~ 2 Å away from the Ti ion; thus the Ti ion and H atoms have all highly positive charges, so that the Ti^{4+} ion does not bind the H_2 molecule or H atoms. A similar situation is found for the $\text{Ti}^{3+}-\text{H}_2$ system at the B3LYP level but not at the MP2 and CCSD(T) levels. At the B3LYP level the ionization energy of H_2 molecule is smaller than the third ionization energy of the titanium atom. Thus, the B3LYP is not reliable for the description of these complexes.

The interaction of Ti^+-H_2 complex was studied experimentally by Bowers' group, who showed that the $4s^13d^2$ configuration is more stable than the $3d^3$ configuration.¹⁶ The experimental ΔE_0 of the ground-state Ti^+-H_2 complex ($4s^13d^2$) was reported to be 7.5/10.0 kcal/mol more stable than the $3d^24s^1$

(ground-state)/ $3d^3$ (excited-state) configuration of Ti^+ . At the B3LYP/W*QZ level, the ground-state configuration of Ti^+ ion is $3d^3$ against the experiment, while at the MP2 and CCSD(T) levels it is $3d^24s^1$ in agreement with the experiment. Thus, the B3LYP/W*QZ binding energy ($-\Delta E_0 = 9.90$ kcal/mol) corresponds to the experimental value of 10.0 kcal/mol with respect to the $3d^3$ excited-state configuration instead of the $3d^24s^1$ ground-state configuration. The MP2/W*QZ and CCSD(T)/W*QZ binding energies ($-\Delta E_0$: 9.89 and 7.70 kcal/mol, respectively) correspond to the experimental value of 7.5 kcal/mol with respect to the $3d^24s^1$ ground-state configuration. Though the MP2/W*QZ value is overestimated in comparison with the experimental value, the CCSD(T)/W*QZ value is in very good agreement with the experiment.

The charge–polarization interaction is a major factor for the binding in these charged complexes. The $\text{Ti}^{3+}-\text{H}_2$ system has the largest polarization interaction among the complexes. The binding energies of the $\text{Ti}^{n+}-\text{H}_2$ complexes ($n = 1-3$) are 9.5, 19.2, and 56.9 kcal/mol, respectively. The main interaction energies arise from the charge(Ti^{n+})–polarization(H_2) interaction. The interaction energies due to H_2 polarization by the positive charge of Ti^{n+} ($n = 1, 2, 3$):

$$\text{polarization energy} = \frac{1}{2}(\text{H}_2\text{polarizability}) \cdot (n \cdot e/r)^2$$

are 8.9, 27.0, and 63.1 kcal/mol, respectively, which are similar to the corresponding total binding energies. In the case of Ti^+-H_2 ($n = 1$), the polarization-driven energy is smaller than the total binding energy, which indicates an additional binding energy for this complexation, i.e., the significant $d-\sigma^*$ backdonation effect. Thus, the Ti–H distance of the Ti^+-H_2 complex is shorter than those of the $\text{Ti}^{2+}-\text{H}_2$ and $\text{Ti}^{3+}-\text{H}_2$ complexes. The ionic radii of Ti^+ , Ti^{2+} , and Ti^{3+} ions are 1.28, 1.00, and 0.81 Å, respectively.²⁹ The ionic radius of Ti^+ is larger than that of Ti^{2+} . However, the Ti^+-H distances are shorter due to the orbital interaction than the $\text{Ti}^{2+}-\text{H}$ distances, as shown in Table 1.

In these charged complexes the H–H stretching frequency (ν_{H_2}) is red-shifted with respect to that of the pure H_2 molecule, as shown in Table 1. The CCSD(T)/W*QZ predicted H–H distances for Ti^+-H_2 , $\text{Ti}^{2+}-\text{H}_2$, and $\text{Ti}^{3+}-\text{H}_2$ are 0.77, 0.77, and 0.84 Å, respectively, with the ν_{H_2} value of 3904, 4037, and 3226 cm^{-1} , respectively. The red-shift is the smallest for the $\text{Ti}^{2+}-\text{H}_2$ complex without $d-\sigma^*$ backdonation, while the red-shift is the largest for the $\text{Ti}^{3+}-\text{H}_2$ complex with strong electrostatic interaction. The asymmetric Ti– H_2 stretching frequency ($\nu_{\text{Ti}-\text{H}_2}^{\text{asym}}$) is the largest in the Ti^+-H_2 complex and the smallest in the $\text{Ti}^{2+}-\text{H}_2$ complex. The CCSD(T)/W*QZ predicted $\nu_{\text{Ti}-\text{H}_2}^{\text{asym}}$ for Ti^+-H_2 , $\text{Ti}^{2+}-\text{H}_2$, and $\text{Ti}^{3+}-\text{H}_2$, are 1057, 783, and 903 cm^{-1} , respectively.

Then, we investigated the structures of dihydrogenated titanium mono- and dications [$\text{Ti}^+(\text{H}_2)_{n=1-7}$ and $\text{Ti}^{2+}(\text{H}_2)_{m=1-9}$]. The geometries were fully optimized at the B3LYP/WTZ and MP2/WTZ levels of theory. An extensive search for the low-lying energy structures including the previously reported ones¹⁶ was made. These structures are shown in Figure 2 for $\text{Ti}^+(\text{H}_2)_n$ and in Figure 3 for $\text{Ti}^{2+}(\text{H}_2)_m$. The dihydrogenated titanium monocations have the spin state of quartet (Q), while the dihydrogenated titanium dications have the triplet state (T). For the notation of each structure, the spin state, the number of H_2 molecules, and the point group of structural symmetry were employed, i.e., Q1-C_{2v} has the spin state of quartet, one H_2 molecule and the point group of C_{2v}. In Figure 3, ‘T’ denotes the triplet state.

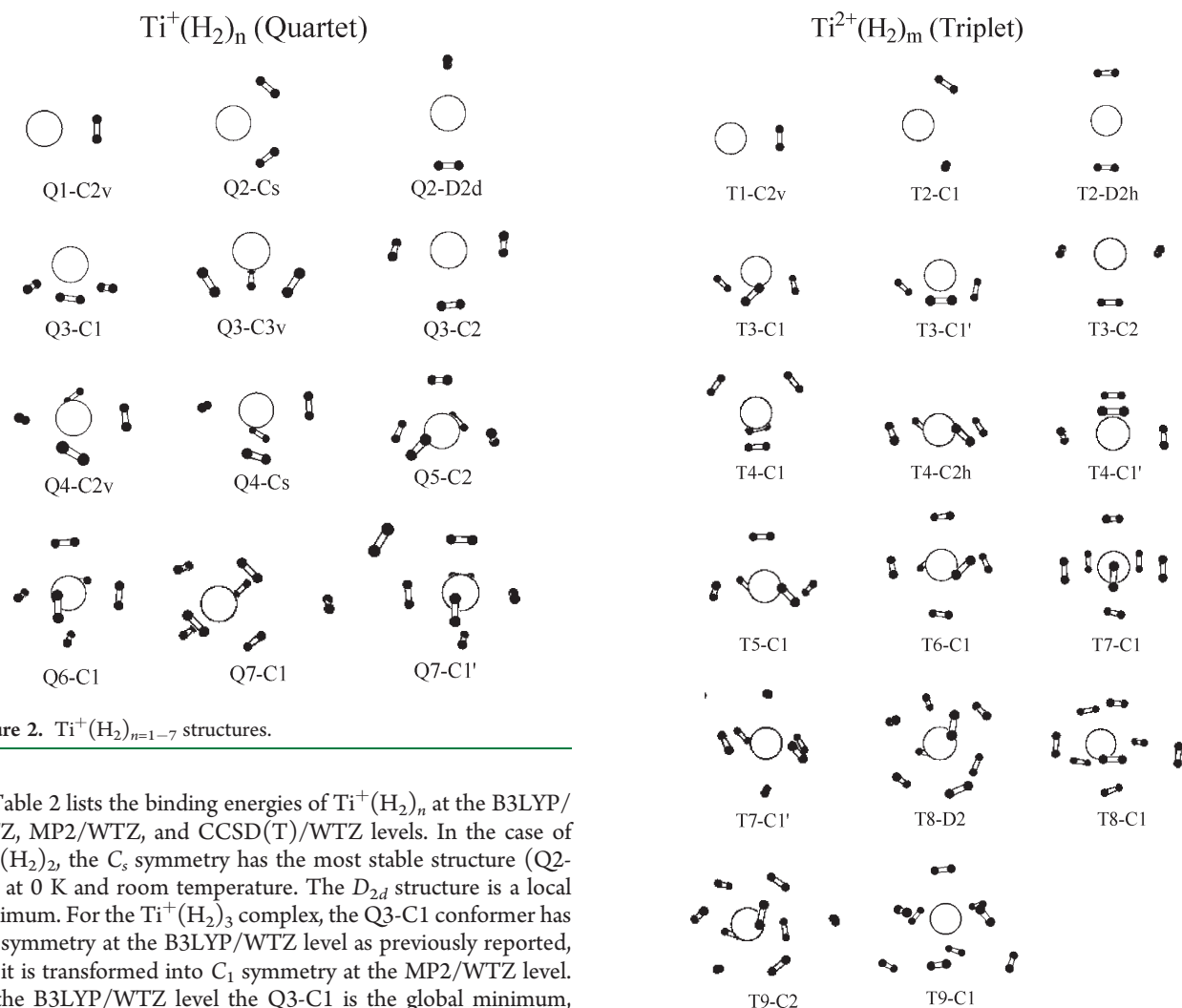


Figure 2. $\text{Ti}^+(\text{H}_2)_{n=1-7}$ structures.

Table 2 lists the binding energies of $\text{Ti}^+(\text{H}_2)_n$ at the B3LYP/WTZ, MP2/WTZ, and CCSD(T)/WTZ levels. In the case of $\text{Ti}^+(\text{H}_2)_2$, the C_s symmetry has the most stable structure (Q2-Cs) at 0 K and room temperature. The D_{2d} structure is a local minimum. For the $\text{Ti}^+(\text{H}_2)_3$ complex, the Q3-C1 conformer has C_{3v} symmetry at the B3LYP/WTZ level as previously reported, but it is transformed into C_1 symmetry at the MP2/WTZ level. At the B3LYP/WTZ level the Q3-C1 is the global minimum, while at the MP2/WTZ and CCSD(T)/WTZ levels the Q3-C3v is the global minimum but almost isoenergetic to the Q3-C1 (within 0.1 kJ/mol at the CCSD(T)/WTZ level). At room temperature the Q3-C1 is the most stable structure at all the levels of calculation. For the $\text{Ti}^+(\text{H}_2)_4$ complex, at the B3LYP/WTZ level the Q4-Cs is the most stable, but at the MP2/WTZ level the Q4-Cs structure transforms into the Q4-C2v structure. In the case of $\text{Ti}^+(\text{H}_2)_5$, the Q5-C2 conformer is the most stable with C_{2v} symmetry.¹² For $\text{Ti}^+(\text{H}_2)_6$, the Q6 has S_4 symmetry at the B3LYP/WTZ level, but C_1 symmetry at the MP2/WTZ level. For $\text{Ti}^+(\text{H}_2)_7$, the Q7-C1 structure which has the coordination 6 + 1 (where 6 and 1 are the numbers of H_2 molecules in the first and second coordination shells, respectively) is the global minimum.

At the CCSD(T)/WTZ level, the $-\Delta E_0$'s of $\text{Ti}^+(\text{H}_2)_{n=1-7}$ are 7.34, 16.16, 24.74, 32.07, 41.99, 48.85, 48.95 kcal/mol, respectively, in good agreement with the experimental values for $n = 1-6$ which are 7.5, 17.2, 26.3, 35.0, 43.2, and 51.9 kcal/mol, respectively. These results indicate that the maximum coordination number of $\text{Ti}^+(\text{H}_2)_n$ is 6 at 0 K. The $-\Delta G^\ddagger$ (room temperature, 1 atm) for $n = 1-7$ is 3.18, 7.93, 9.42, 11.59, 14.27, 14.15, and 11.09 kcal/mol, respectively, indicating the optimal coordination number of 5–6 at room temperature, as shown in Figure 4. We estimated the approximate complete basis set (aCBS) limit value of the binding energies at the CCSD(T) level, using the extrapolation scheme by utilizing that the electron correlation

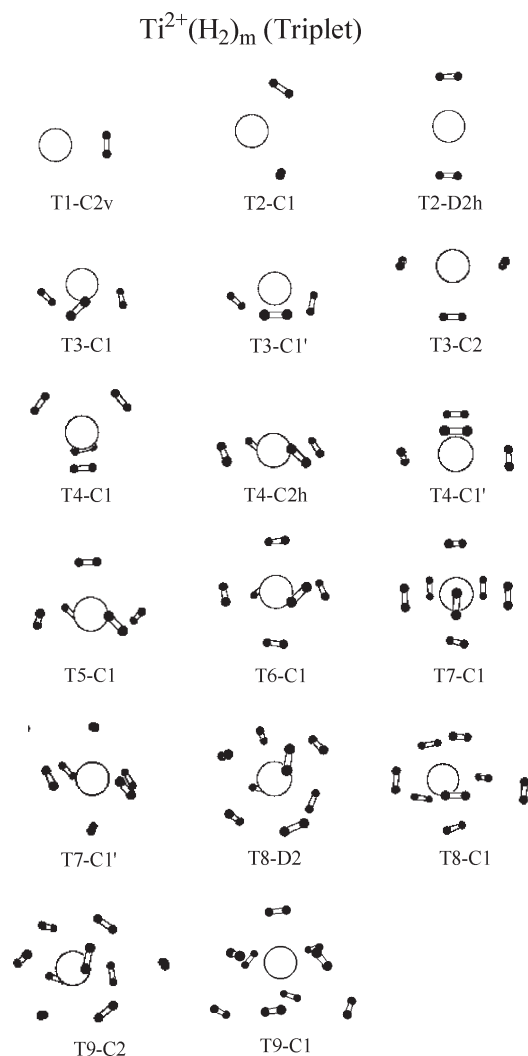


Figure 3. $\text{Ti}^{2+}(\text{H}_2)_{m=1-9}$ structures.

error is proportional to N^{-3} for the (aug)-cc-pVNZ basis set,³⁰ where we used $N = 2, 3$, while N is generally required to use more than 3 and 4.³¹ The estimated CCSD(T)/aCBS $-\Delta E_0$'s of $\text{Ti}^+(\text{H}_2)_1$ and $\text{Ti}^+(\text{H}_2)_6$ are 7.5 and 49.1 kcal/mol. For the Ti^+-H_2 complex the estimated CCSD(T)/aCBS binding energy (ΔE_0) is closer to the experimental value (-7.5 kcal/mol) than the CCSD(T)/WTZ value (-7.3 kcal/mol).

The B3LYP/WTZ, MP2/WTZ, and CCSD(T)/WTZ binding energies of $\text{Ti}^{2+}(\text{H}_2)_{m=1-9}$ are listed in Table 3. The $\text{Ti}^{2+}(\text{H}_2)$ complex has C_{2v} symmetry and the $-\Delta E_0$ is 16.20 kcal/mol at the CCSD(T)/WTZ level, 8.86 kcal/mol larger than that (7.34 kcal/mol) of $\text{Ti}^+(\text{H}_2)$. For the $\text{Ti}^{2+}(\text{H}_2)_2$ complex the T2-C1 structure is the global minimum. The T2-D2h structure has one imaginary frequency at the B3LYP/WTZ level, while T2-D2d has one imaginary frequency at the MP2/WTZ level. In the case of $\text{Ti}^{2+}(\text{H}_2)_3$, the T3-C1 structure is the global minimum, but the T3-C1' structure is nearly iso-energetic within 0.4 kcal/mol at the CCSD(T)/WTZ level. The $\text{Ti}^{2+}(\text{H}_2)_{4-7}$ complexes have C_1 symmetry. The $\text{Ti}^{2+}(\text{H}_2)_8$ complex has the T8-D2 structure (8 + 0). The T8-C1 structure (7 + 1) is slightly less stable. For $\text{Ti}^{2+}(\text{H}_2)_9$, the most stable structure is T9-C2 (8 + 1) at 0 K and T9-C1 (7 + 2) at room temperature, while the (9 + 0) structure is not stable. At the CCSD(T)/WTZ level, the $-\Delta E_0$'s of the

Table 2. Binding Energies (kcal/mol) of $\text{Ti}^+(\text{H}_2)_{n=1-7}$ ^a

$n\text{H}_2$	conformer (sym.) ^b	$(n_1 + n_2)^c$	B3LYP/WTZ				MP2/WTZ				CCSD(T)/WTZ ^d			
			$-\Delta E_e$	$-\Delta E_0$	$-\Delta H_r$	$-\Delta G_r$	$-\Delta E_e$	$-\Delta E_0$	$-\Delta H_r$	$-\Delta G_r$	$-\Delta E_e$	$-\Delta E_0$	$-\Delta H_r$	$-\Delta G_r$
1	Q1-C2v(C_{2v})	(1 + 0)	12.15	10.47	11.57	6.35	12.28	10.39	11.50	6.24	9.20	7.34	8.44	3.18
2	Q2-Cs(C_s)	(2 + 0)	23.81	19.53	21.54	10.75	23.18	18.54	20.51	10.33	20.79	16.16	18.12	7.93
	Q2-D2d(D_{2d})	(2 + 0)	23.16	18.66	20.85	8.23	22.79	17.82	20.06	7.32	19.91	14.94	17.18	4.45
3	Q3-C1(C_{3v} , C_1)	(3 + 0)	35.17	27.88	31.07	12.20	34.76	26.66	30.05	11.77	32.84	24.74	28.13	9.42
	Q3-C3v(C_{3v})	(3 + 0)	35.07	27.77	30.96	12.05	34.89	26.73	30.12	10.80	32.91	24.74	28.13	8.82
	Q3-C2(C_2)	(3 + 0)	34.52	27.26	30.42	11.87	—	—	—	—	—	—	—	—
4	Q4-C2v(C_{2v})	(4 + 0)	44.63	34.37	38.68	12.64	45.03	34.13	38.37	13.67	42.97	32.07	36.30	11.59
	Q4-Cs(C_s)	(4 + 0)	45.42	35.11	39.45	13.73	—	—	—	—	—	—	—	—
5	Q5-C2(C_2)	(5 + 0)	55.20	42.05	47.40	14.16	54.29	39.69	45.36	11.98	56.60	41.99	47.66	14.27
6	Q6-C1(S_6 , C_1)	(6 + 0)	64.46	48.61	54.83	15.50	65.54	46.93	54.17	12.63	67.47	48.85	56.09	14.15
7	Q7-C1(C_1)	(6 + 1)	64.86	48.16	54.21	12.41	66.36	46.95	53.99	9.51	68.36	48.95	56.00	11.09
	Q7-C1'(C_1)	(6 + 1)	64.66	47.89	54.00	11.94	63.91	44.53	51.56	6.65	63.22	43.83	50.86	5.54

^a Binding energy: $E[\text{Ti}^+(\text{H}_2)_n] - E[\text{Ti}^+] - nE[\text{H}_2]$. ^b The conformational symmetries at the B3LYP/WTZ and MP2/WTZ levels are presented in parentheses, respectively; only one group symmetry is given when both symmetries are the same. ^c Both n_1 and n_2 are the numbers of the H_2 molecules in the first and the second solvation shells, respectively. ^d The CCSD(T) data using the WTZ basis set at the MP2/WTZ geometries. The free energy changes of chiral conformers were corrected for chirality by $-RT \ln 2$ (R , gas constant; T , temperature). The experimental $-\Delta E_0$ for $n = 1-6$ is 7.5, 17.2, 26.5, 35.0, 43.2, and 51.9 kcal/mol for the ground-state $4s^1 3d^2$ configuration (ref 16).

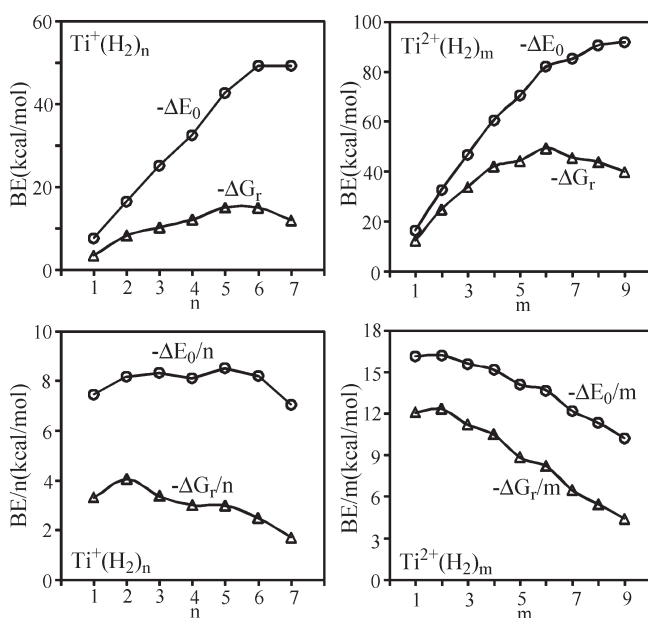


Figure 4. CCSD(T)/WTZ binding energies of the lowest-energy structures of $\text{Ti}^+(\text{H}_2)_n$ and $\text{Ti}^{2+}(\text{H}_2)_m$.

lowest-energy structures of $\text{Ti}^{2+}(\text{H}_2)_{m=1-9}$ are 16.20, 32.53, 46.94, 60.80, 70.70, 82.19, 85.09, 89.60, and 90.73 kcal/mol, respectively. Here, the case for $n = 9$ shows the coordination $8 + 1$, while all other lowest-energy structures for $n < 9$ show no secondary coordination. This indicates that the maximum coordination number of $\text{Ti}^{2+}(\text{H}_2)_m$ is 8 at 0 K. The $-\Delta G_r$ for $n = 1-9$ is 12.17, 24.40, (33.80), 41.75, 44.02, 49.00, (45.20), 42.50, and 38.81 kcal/mol, respectively, which indicates the optimal coordination number of 6 at room temperature (Figure 4). Here, the values in parentheses mean that the entropy effect of their two nearly isoenergetic conformers is taken into account.

As shown in Figure 4, for $\text{Ti}^+(\text{H}_2)_{n=1-7}$ and $\text{Ti}^{2+}(\text{H}_2)_{m=1-9}$ at room temperature, the cases of $n = 2$ and $m = 2$ have the largest $-\Delta G_r/n$ and $-\Delta G_r/m$ values (3.97 and 12.20 kcal/mol), respectively. Even though n and m increase, the values of $-\Delta G_r/n$ and $-\Delta G_r/m$ decrease, but the $-\Delta G_r/m$ (4.30 kcal/mol) of $\text{Ti}^{2+}(\text{H}_2)_9$ is larger than $-\Delta G_r/n$ (3.97 kcal/mol) of $\text{Ti}^+(\text{H}_2)_2$. The ionic radius (1.28 Å) of Ti^+ is larger than that (1.00 Å) of Ti^{2+} . However, the Ti^+-H distances are shorter due to the orbital interaction than the $\text{Ti}^{2+}-\text{H}$ distances as shown in Table 1. That is, the solvation sphere of Ti^{2+} is larger than that of Ti^+ ion. The Ti^{2+} ion has more H_2 molecules in the first solvation shell than the Ti^+ ion. Due to the stronger electrostatic interaction by the Ti^{2+} ion, $\text{Ti}^{2+}(\text{H}_2)_m$ have stronger binding energies with a larger coordination number than $\text{Ti}^+(\text{H}_2)_n$.

Table 4 shows the MP2/WTZ NBO charges and selected geometrical parameters of $\text{Ti}^+(\text{H}_2)_{n=1-7}$ and $\text{Ti}^{2+}(\text{H}_2)_{m=1-9}$. For the $\text{Ti}^+(\text{H}_2)_{n=1-7}$ complexes the Q1-C2v and Q2-Cs conformers show electron backdonation from the Ti 3d orbital to H_2 σ^* orbital,¹⁶ and so the NBO charges of Ti ion are larger than +1. For the larger size clusters the electron transfers from the H_2 molecules to the Ti ion. The Ti-H distances for the H_2 molecules in the first coordination shell are around 2.0 Å, while those for the H_2 molecules in the second coordination shell are over 4.0 Å. The H-H distances of H_2 molecules in the complexes are elongated in comparison with that of the pure H_2 molecule. For the $\text{Ti}^{2+}(\text{H}_2)_m$ complexes, the electron transfers from the H_2 molecules to the Ti ion, without the d- σ^* backdonation. The Ti-H distances of $\text{Ti}^{2+}(\text{H}_2)_m$ are longer than those of $\text{Ti}^+(\text{H}_2)_n$ involved in the d- σ^* backdonation, while the H-H distances of H_2 molecules in $\text{Ti}^{2+}(\text{H}_2)_m$ are shorter than those in $\text{Ti}^+(\text{H}_2)_n$. Figure 5 shows the MP2/WTZ NBO charges of titanium ions in the ΔE_0 -based lowest-energy structures of $\text{Ti}^+(\text{H}_2)_n$ and $\text{Ti}^{2+}(\text{H}_2)_m$. The charge transfer from hydrogen molecules to the titanium ion increases up to $n = 6$ and $m = 7$ where the charge transfers are almost saturated. The Ti^+ ion accepts charge by -0.58 au in Q6-C1, while the Ti^{2+} ion accepts charge by -0.63 au in T6-C1. In $\text{Ti}^+(\text{H}_2)_n$ the charge ($q_{\text{NBO}}^{\text{Ti}}$)

Table 3. Binding Energies (kcal/mol) of $\text{Ti}^{2+}(\text{H}_2)_{m=1-9}^a$

$m\text{H}_2$	conformer (sym.)	$(n_1 + n_2)$	B3LYP/WTZ				MP2/WTZ				CCSD(T)/WTZ			
			$-\Delta E_e$	$-\Delta E_0$	$-\Delta H_r$	$-\Delta G_r$	$-\Delta E_e$	$-\Delta E_0$	$-\Delta H_r$	$-\Delta G_r$	$-\Delta E_e$	$-\Delta E_0$	$-\Delta H_r$	$-\Delta G_r$
1	T1-C2v(C_{2v})	(1 + 0)	20.37	18.92	19.98	14.89	17.49	15.85	16.92	11.81	17.83	16.20	17.28	12.17
2	T2-C1(C_1)	(2 + 0)	40.99	37.49	39.13	30.22	35.84	31.83	33.60	24.10	36.54	32.53	34.32	24.40
	T2-D2h(D_{2h}) ^b	(2 + 0)	39.12	35.38	37.79	24.84	35.70	31.74	33.49	22.87	36.35	32.41	34.15	23.54
3	T3-C1(C_1)	(3 + 0)	59.05	53.19	55.65	40.05	52.32	45.84	48.39	32.70	53.42	46.94	49.47	33.39
	T3-C1'(C_1)	(3 + 0)	57.96	51.47	54.39	37.17	52.31	45.81	48.39	32.65	53.37	46.87	49.45	33.32
	T3-C2(C_{2v})	(3 + 0)	53.68	47.10	49.98	32.49	—	—	—	—	—	—	—	—
4	T4-C1(C_1)	(4 + 0)	75.67	67.58	70.71	49.39	68.22	59.29	62.63	40.67	69.72	60.80	64.13	41.75
	T4-C2h(C_{2h})	(4 + 0)	73.67	65.03	68.59	44.46	66.40	56.70	60.65	35.53	67.97	58.27	62.21	37.09
	T4-C1b(C_1)	(4 + 0)	70.93	61.55	65.44	42.31	63.92	53.45	57.69	32.98	66.40	55.93	60.16	35.04
5	T5-C1(C_{3v})	(5 + 0)	88.81	77.11	81.95	50.13	81.32	68.94	73.81	42.68	83.08	70.70	75.57	44.02
6	T6-C1(C_1)	(6 + 0)	102.53	87.86	93.87	54.65	94.90	79.64	85.63	46.86	97.44	82.19	88.17	49.00
7	T7-C1(C_{2v})	(7 + 0)	110.10	92.41	99.57	52.69	103.34	81.83	90.35	41.99	106.60	85.09	93.59	44.81
	T7-C1'(C_1)	(6 + 1)	104.95	88.68	95.04	50.62	97.85	80.86	87.31	42.66	100.50	83.53	89.96	44.91
8	T8-D2(D_2)	(8 + 0)	113.36	93.77	101.63	47.44	107.86	86.39	95.04	39.28	111.07	89.60	98.23	42.50
	T8-C1(C_1)	(7 + 1)	112.39	93.25	100.68	50.30	105.75	84.38	92.74	39.11	109.37	88.00	96.37	42.33
9	T9-C2(C_2)	(8 + 1)	115.21	94.09	102.25	43.19	110.53	87.41	96.48	34.82	113.86	90.73	99.81	38.15
	T9-C1(C_1)	(7 + 2)	114.66	93.91	101.73	44.60	108.78	86.33	94.84	35.69	112.33	89.87	98.37	38.81

^a See the footnote of Table 2. ^b Though this structure is a transition state at B3LYP/WTZ, it is a minimum at MP2/WTZ.

Table 4. MP2/WTZ NBO Charges (q_{NBO}) and Geometrical Parameters (Å) of $\text{Ti}^+(\text{H}_2)_n$ and $\text{Ti}^{2+}(\text{H}_2)_m$ ^a

$\text{Ti}^+(\text{H}_2)_n$				$\text{Ti}^{2+}(\text{H}_2)_m$			
conf.	$q_{\text{NBO}}^{\text{Ti}}$	$r_{\text{Ti-H}}^{\text{av}}$	$r_{\text{HH}}^{\text{av}}$	conf.	$q_{\text{NBO}}^{\text{Ti}}$	$r_{\text{Ti-H}}^{\text{av}}$	$r_{\text{HH}}^{\text{av}}$
Q1-C2v	1.03	1.983	0.765	T1-C2v	1.98	2.151	0.765
Q2-Cs	1.02	1.983	0.767	T2-C1	1.91	2.135	0.763
Q3-C1	0.95	1.976	0.769	T2-D2h	1.92	2.148	0.762
Q3-C3v	0.95	1.976	0.769	T3-C1	1.82	2.139	0.761
Q4-C2v	0.83	2.013	0.765	T3-C1'	1.82	2.139	0.761
Q5-C2	0.66	2.001	0.767	T4-C1	1.70	2.141	0.760
Q6-C1	0.42	2.002	0.766	T5-C1	1.56	2.137	0.760
Q7-C1	0.42	2.002	0.766	T6-C1	1.37	2.134	0.759
		(4.087)	(0.739)	T7-C1	1.14	2.129	0.759
				T8-D2	1.10	2.224	0.753
				T9-C2	1.08	2.223	0.753
						(3.600)	(0.742)
				T9-C1	1.26	2.150	0.757
						(3.481)	(0.743)

^a $q_{\text{NBO}}^{\text{Ti}}$ is the NBO charge of the Ti ion; $r_{\text{Ti-H}}^{\text{av}}$ is the average Ti-H distance; and $r_{\text{HH}}^{\text{av}}$ is the average H-H distance of the first-shell H_2 molecules. The data in parentheses are of the secondary shell H_2 molecules. The H-H distance of the pure H_2 is 0.738 Å at the MP2/WTZ level.

converges to +0.42 au as the number of H_2 molecules increases up to $n = 6$. In $\text{Ti}^{2+}(\text{H}_2)_m$ the charge converges to +1.14 au as the number of H_2 molecules increases up to $m = 7$, as shown in Figure 5. At the MP2/WTZ level the average Ti^+-H distances in Q6-C1 and Q7-C1 are 2.002 Å, while the average $\text{Ti}^{2+}-\text{H}$ distances in T6-C1 is 2.134 Å. The shortest intermolecular distances between the nearest-neighboring H_2 centers in the first solvation shells of Q6-C1 and Q7-C1 are 2.65 Å. The

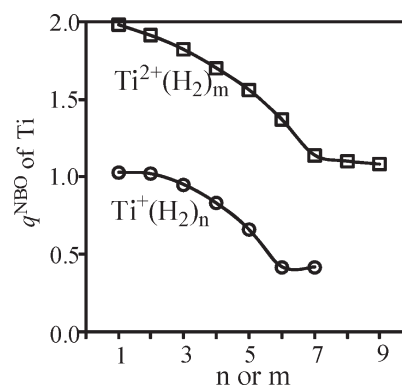


Figure 5. MP2/WTZ NBO charge (q^{NBO}) of the titanium ion in the ΔE_0 -based lowest-energy structures of $\text{Ti}^+(\text{H}_2)_n$ and $\text{Ti}^{2+}(\text{H}_2)_m$.

shortest distances between the nearest-neighboring H_2 centers in the first solvation shells of T6-C1, T7-C1, T8-D2, and T9-C2 are 2.86, 2.48, 2.53, and 2.53 Å, respectively. The shortest H_2-H_2 distances for hepta-, octa-, and nonacoordinated systems are already small enough to accommodate an extra H_2 .

The vibrational frequencies of $\text{Ti}^+(\text{H}_2)_n$ were calculated at the B3LYP/WTZ and MP2/WTZ levels. Table 5 lists the MP2 (B3LYP) frequencies scaled by 0.975 (0.966) to match the frequency of pure H_2 with respect to the experimental value (4401 cm^{-1})²⁸ and their IR intensities. The frequencies and their intensities of $\text{Ti}^{2+}(\text{H}_2)_m$ are listed in Table 6. Generally, in the case of electrostatic energy driven interaction systems the H-H stretching frequencies increase and the asymmetric Ti ion- H_2 stretching frequencies decrease with the increasing number of H_2 molecules. However, this trend does not appear due to the complex orbital interactions, especially in the $\text{Ti}^+(\text{H}_2)_n$ cases. The $\text{Ti}^+(\text{H}_2)_n$ complexes have orbital interactions due to the $d-\sigma^*$ backdonation in addition to weak electrostatic interactions,

Table 5. Selected Frequencies and IR Intensities in Subscripts (10 km/mol) of $\text{Ti}^+(\text{H}_2)_n$ at the MP2/WTZ (B3LYP/WTZ) level^a

<i>n</i>	conformer	H–H stretch	asym. $\text{Ti}^+ - \text{H}_2$ stretch in the first shell
1	Q1-C2v	3910 ₁₃ (3841 ₁₃)	1051 (1027)
2	Q2-Cs	3884 ₆ 3900 ₁₈ (3859 ₁₈ 3863 ₅)	1070 1083 (1033 1037)
3	Q3-C1	3855 ₃ 3869 ₁₆ 3870 ₁₆ (3875 ₁₃ 3875 ₁₃ 3881 ₂)	1078 1102 1102 (1031 1043 1043)
	Q3-C3v	3860 ₁ 3866 ₂₀ 3866 ₂₀ (3878 ₁₅ 3878 ₁₅ 3886 ₁)	1100 1100 1115 (1037 1037 1045)
4	Q4-C2v	3930 ₀ 3935 ₁₈ 3935 ₁₈ 3948 ₀ (3937 ₂₈ 3938 ₂₃ 3938 ₀ 3941 ₀)	1033 1057 1057 1059 (1001 1003 1007 1007)
5	Q5-C2	3864 ₅ 3915 ₂ 3920 ₁₅ 3926 ₁₅ 3932 ₀ (3920 ₃ 3951 ₀ 3952 ₁₃ 3954 ₁₁ 3956 ₁)	1058 1068 1071 1073 1123 (999 1003 1004 1006 1035)
6	Q6-C1	3910 ₀ 3915 ₅ 3916 ₅ 3924 ₇ 3928 ₅ 3929 ₅ (3968 ₀ 3968 ₁ 3973 ₈ 3973 ₈ 3973 ₈ 3980 ₀)	1089 1089 1090 1090 1092 1093 (994 996 996 999 1000 1000)
7	Q7-C1	3910 ₁ 3915 ₆ 3919 ₂ 3923 ₇ 3929 ₆ 3934 ₄ 4380 ₁ (3966 ₂ 3968 ₁ 3972 ₇ 3973 ₈ 3976 ₆ 3980 ₁ 4384 ₁)	1088 1089 1091 1092 1092 1094 (996 997 998 1001 1001 1002)

^a MP2 (B3LYP) frequencies are scaled by 0.975 (0.966) to match the frequency of pure H_2 with respect to the experimental value (4401 cm^{-1}) from ref 28. The intensities of $\text{Ti}^+ - \text{H}_2$ stretches are omitted due to their small values.

Table 6. Selected Frequencies and IR Intensities in Subscripts (10 km/mol) of $\text{Ti}^{2+}(\text{H}_2)_m$ at the MP2/WTZ (B3LYP/WTZ) Level^a

<i>m</i>	conformer	H–H stretch	asym. $\text{Ti}^{2+} - \text{H}_2$ stretch
1	T1-C2v	4033 ₁₅ (3965 ₁₉)	784 (760)
2	T2-C1	4058 ₂₁ 4065 ₉ (4003 ₂₆ 4014 ₉)	794 797 (765 771)
3	T3-C1	4079 ₁₈ 4079 ₁₈ 4087 ₃ (4028 ₂₃ 4028 ₂₃ 4042 ₃)	799 802 804 (775 776 778)
	T3-C1'	4078 ₁₈ 4078 ₁₈ 4086 ₃ (4023 ₁₃ 4023 ₂₀ 4033 ₇)	796 797 804 (791 795 797)
4	T4-C1	4090 ₁₇ 4091 ₁₅ 4091 ₁₅ 4100 ₀ (4055 ₁₉ 4055 ₁₇ 4056 ₁₇ 4070 ₀)	797 803 807 809 (776 778 782 786)
5	T5-C1	4086 ₉ 4095 ₁₈ 4097 ₁₁ 4098 ₂ 4100 ₅ (4055 ₁₀ 4068 ₂₁ 4071 ₁₈ 4073 ₁ 4079 ₁)	792 812 814 826 833 (787 793 797 799 821)
6	T6-C1	4096 ₁₃ 4096 ₅ 4096 ₁₈ 4098 ₀ 4098 ₀ 4101 ₁₄ (4078 ₁₉ 4078 ₁₉ 4082 ₀ 4082 ₁ 4082 ₁₄ 4086 ₀)	786 813 818 818 827 827 (796 806 806 807 811 812)
7	T7-C1	4074 ₀ 4077 ₁₂ 4110 ₀ 4110 ₀ 4114 ₁₉ 4115 ₁₉ 4120 ₀ (4082 ₁₅ 4083 ₀ 4109 ₂ 4111 ₉ 4114 ₁₇ 4115 ₁₃ 4123 ₂)	821 822 846 859 860 861 862 (774 775 778 781 792 810 811)
8	T8-D2	4189 ₀ 4191 ₂₁ 4192 ₁₇ 4192 ₂₀ 4193 ₂ 4193 ₁ 4193 ₄ 4195 ₀ (4179 ₂₄ 4179 ₂₄ 4179 ₂₃ 4179 ₀ 4181 ₁ 4182 ₀ 4182 ₁ 4190 ₀)	687 691 691 696 701 704 704 707 (695 695 697 698 702 704 705 706)
9	T9-C2	4181 ₉ 4183 ₉ 4193 ₈ 4194 ₁₂ 4196 ₃ 4197 ₉ 4197 ₉ 4199 ₄ 4336 ₄ (4169 ₁₃ 4172 ₁₃ 4179 ₈ 4181 ₂ 4185 ₁₈ 4185 ₁₂ 4189 ₅ 4194 ₀ 4314 ₅)	684 690 695 699 704 707 729 739 (694 696 702 703 708 709 734 736)
	T9-C1	4100 ₀ 4100 ₁₅ 4133 ₀ 4134 ₁₁ 4138 ₉ 4140 ₇ 4143 ₁₀ 4323 ₇ 4323 ₄ (4070 ₁₆ 4072 ₀ 4100 ₆ 4106 ₁₅ 4116 ₄ 4119 ₁₅ 4124 ₃ 4328 ₉ 4328 ₅)	774 793 800 808 819 836 838 (772 782 790 825 830 839 840)

^a See the footnote of Table 5.

whereas the $\text{Ti}^{2+}(\text{H}_2)_m$ complexes have relatively strong electrostatic interactions. Owing to the $d-\sigma^*$ backdonation, the Ti ion–H distances of $\text{Ti}^+(\text{H}_2)_n$ are shorter than those of $\text{Ti}^{2+}(\text{H}_2)_m$. Since the H–H distances of H_2 molecules in $\text{Ti}^+(\text{H}_2)_n$ complexes are longer than those in $\text{Ti}^{2+}(\text{H}_2)_m$ due to the $d-\sigma^*$ backdonation, the H–H stretching frequencies of $\text{Ti}^+(\text{H}_2)_n$ are smaller (more shifted in comparison with the frequency of pure H_2 molecule) than those of $\text{Ti}^{2+}(\text{H}_2)_m$. On the other hand, the asymmetric $\text{Ti}^{2+} - \text{H}_2$ stretching frequencies of $\text{Ti}^{2+}(\text{H}_2)_m$ are smaller than the asymmetric $\text{Ti}^+ - \text{H}_2$ stretching frequencies of $\text{Ti}^+(\text{H}_2)_n$. The B3LYP/WTZ IR spectra for the H–H stretching frequencies of the ΔE_0 -based lowest-energy structures of $\text{Ti}^+(\text{H}_2)_n$ and $\text{Ti}^{2+}(\text{H}_2)_m$ are shown in Figure 6. The red shifts of H–H stretching frequencies decrease with increasing number of hydrogen molecules in both $\text{Ti}^+(\text{H}_2)_n$ and $\text{Ti}^{2+}(\text{H}_2)_m$. The red shift becomes minimal at $n = 6$ in the $\text{Ti}^+(\text{H}_2)_n$ complexes and at $m = 8$ in the $\text{Ti}^{2+}(\text{H}_2)_m$ complexes.

At 373 K, ΔG_{373} 's of Q6-C1, T6-C1, T7-C1, T7-C1', and T8-D2 were calculated to be -20.82 , -39.67 , -33.03 , -34.11 , and -28.49 kcal/mol, based on the CCSD(T) energies with the MP2 thermal energy corrections. These binding free energies are most suitable to store and release hydrogen molecules. As Ti^{n+} cations should be neutralized with anions in real hydrogen storage materials, researchers have been trying to design optimal systems. Transition metals have the tendency to aggregate among themselves, resulting in the reduction of the hydrogen storage capacity. Thus, to stabilize the dispersed transition metals, they can be embedded in frameworks or decorated on the substrate. To bind hydrogen directly, the Ti cations should be in the (I–III) oxidation states, not in the (IV) oxidation state. Hypothetical studies of Ti-decorated bulkyballs ($\text{Ti}(\text{I})_{12}\text{B}_{24}\text{C}_{36}$, 8.6 wt % H_2),³² $\text{Ti}(\text{I–III})$ –graphene oxide (4.9 wt % H_2),³³ $\text{Ti}(\text{I})$ –nanotubes (5–8 wt % H_2),³⁴ $\text{Ti}(\text{I})$ –ethane-diol (13 wt % H_2),³⁵ and Ti-substituted boranes ($\text{Ti}(\text{II})_n - \text{B}_m\text{H}_m$, maximum 10.0 wt % H_2)³⁶ have been carried out. The

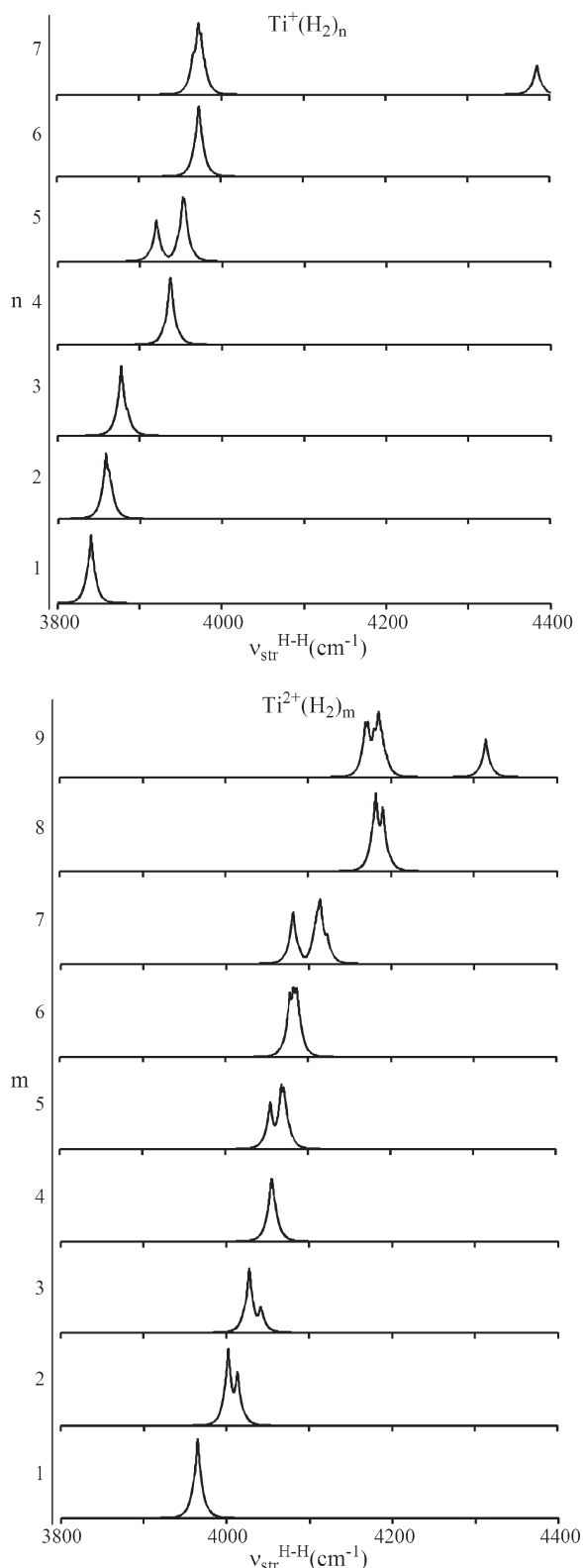


Figure 6. B3LYP/WTZ vibrational frequencies for the H–H stretching modes of the ΔE_0 -based lowest-energy structures of $\text{Ti}^+(\text{H}_2)_n$ and $\text{Ti}^{2+}(\text{H}_2)_m$.

$\text{Ti}^{2+}(\text{H}_2)_8$ complexation (25.2 wt %) would help design effective hydrogen storage materials by Ti-decoration on or Ti-dispersion in graphene derivatives, organic π systems with

strongly electronegative substituents, or polymer–graphene hybrid materials.

IV. CONCLUDING REMARKS

The neutral and multiply charged titanium– H_2 systems ($\text{Ti}-\text{H}_2$, Ti^+-H_2 , $\text{Ti}^{2+}-\text{H}_2$, $\text{Ti}^{3+}-\text{H}_2$ and $\text{Ti}^{4+}-\text{H}_2$) were calculated using the B3LYP, MP2 and CCSD(T) methods with the W^*QZ basis set. The neutral $\text{Ti}-\text{H}_2$ system has the covalent-bonded dihydride configuration with strong $d-\sigma^*$ backdonation. The Ti^+-H_2 , $\text{Ti}^{2+}-\text{H}_2$ and $\text{Ti}^{3+}-\text{H}_2$ complexes have the non-covalent-bonded dihydrogen configurations with the electrostatic interaction, resulting in electron donation from the H_2 σ orbital to the metal which stabilizes the ion charge. Especially for the Ti^+-H_2 complex the empty $4s$ orbital plays a crucial role in the stability. The B3LYP calculation fails to describe the binding between a titanium trication (Ti^{3+}) and a hydrogen molecule. The Ti^{4+} ion does not bind a H_2 molecule. The predicted binding energy of the Ti^+-H_2 complex is in good agreement with the experimental value. Among the Ti^+-H_2 , $\text{Ti}^{2+}-\text{H}_2$, and $\text{Ti}^{3+}-\text{H}_2$ complexes, the Ti^+-H_2 system has a characteristic bonding orbital with the $d-\sigma^*$ backdonation which leads to the shortest Ti–H distance and the highest Ti– H_2 asymmetric stretching frequency; the $\text{Ti}^{3+}-\text{H}_2$ system has the strong electrostatic interaction; the $\text{Ti}^{2+}-\text{H}_2$ system has the longest Ti–H distance, resulting in the smallest red shift in the H–H stretching frequency and the smallest Ti– H_2 asymmetric stretching frequency.

Then the coordination structures of $\text{Ti}^+(\text{H}_2)_{n=1-7}$ and $\text{Ti}^{2+}(\text{H}_2)_{m=1-9}$ were studied at the B3LYP/WTZ, MP2/WTZ, and CCSD(T)/WTZ levels of theory. At low temperatures the most stable structures of $\text{Ti}^+(\text{H}_2)_{n=1-7}$ are Q1-C2v (1 + 0), Q2-Cs (2 + 0), Q3-C3v (3 + 0), Q4-C2v (4 + 0), Q5-C2 (5 + 0), Q6-C1 (6 + 0), and Q7-C1 (6 + 1), while at room temperature the most stable structures are Q1-C2v (1 + 0), Q2-Cs (2 + 0), Q3-C1 (3 + 0), Q4-C2v (4 + 0), Q5-C2 (5 + 0), Q6-C1 (6 + 0), and Q7-C1b (6 + 1). For the $\text{Ti}^{2+}(\text{H}_2)_{m=1-9}$ complexes, T1-C2v (1 + 0), T2-C1 (2 + 0), T3-C1 (3 + 0), T4-C1 (4 + 0), T5-C1 (5 + 0), T6-C1 (6 + 0), T7-C1 (7 + 0), T8-D2 (8 + 0), and T9-C2 (8 + 1) are the most stable structures at low temperatures, while T1-C2v (1 + 0), T2-C1 (2 + 0), T3-C1/T3-C1 (3 + 0), T4-C1 (4 + 0), T5-C1 (5 + 0), T6-C1 (6 + 0), T7-C1/T7-C1' (7 + 0), T8-D2 (8 + 0) and T9-C2 (7 + 2) are the most stable structures at room temperature. The $\text{Ti}^+(\text{H}_2)_n$ complexes have orbital interactions due to the $d-\sigma^*$ electron backdonation in addition to weak electrostatic interactions, while the $\text{Ti}^{2+}(\text{H}_2)_m$ complexes have relatively strong electrostatic interactions. Thus, the binding energies of $\text{Ti}^{2+}(\text{H}_2)_m$ complexes are larger than those of $\text{Ti}^+(\text{H}_2)_n$. However, the Ti–H distances of $\text{Ti}^+(\text{H}_2)_n$ are shorter than those of $\text{Ti}^{2+}(\text{H}_2)_m$ and the H–H distances of H_2 molecules in $\text{Ti}^+(\text{H}_2)_n$ complexes are longer than those in $\text{Ti}^{2+}(\text{H}_2)_m$. Thus, the H–H stretching frequencies of $\text{Ti}^+(\text{H}_2)_n$ are smaller than those of $\text{Ti}^{2+}(\text{H}_2)_m$, whereas the asymmetric $\text{Ti}^{2+}-\text{H}_2$ stretching frequencies of $\text{Ti}^{2+}(\text{H}_2)_m$ are smaller than the asymmetric Ti^+-H_2 stretching frequencies of $\text{Ti}^+(\text{H}_2)_n$. For $\text{Ti}^+(\text{H}_2)_n$ and $\text{Ti}^{2+}(\text{H}_2)_m$ it is possible to have up to the hexacoordination and octacoordination at very low temperatures, respectively, while they favor penta- to hexacoordination and hexacoordination, respectively, at room temperature and 1 atm. This coordination structure would be important information for the design of hydrogen storage material of Ti complexes. Indeed, many Ti-decorated systems have been studied for their H_2 storage properties.^{10,32–36} From this study, we find that the titanium atoms

need to have positive charges (preferentially doubly charged state which allows the primary coordination number of 6 at ambient conditions and up to 8 at very low temperatures near 0 K) for better H₂ storage.

AUTHOR INFORMATION

Corresponding Author

*E-mail: kim@postech.ac.kr.

ACKNOWLEDGMENT

This work is affectionately dedicated to Professor Bernd Brutschy, an outstanding scientist, on the occasion of his 65th birthday. This work was supported by NRF (WCU: R32-2008-000-10180-0, BK21, National honor scientist program) and KISTI (KSC-2008-K08-0002).

REFERENCES

- (1) Coontz, R.; Hanson, B. *Science* **2004**, *305*, 957.
- (2) Orimo, S.-I.; Nakamori, Y.; Eliseo, J. R.; Züttel, A.; Jensen, C. M. *Chem. Rev.* **2007**, *107*, 4111–4132.
- (3) Marder, C. T. B. *Angew. Chem., Int. Ed.* **2007**, *46*, 8116–8118.
- (4) (a) Xiong, Z.; Yong, C. K.; Wu, G.; Chen, P.; Shaw, W.; Karkamkar, A.; Autrey, T.; Jones, M. O.; Johnson, S. R.; Edwards, P. P.; David, W. I. F. *Nat. Mater.* **2008**, *7*, 138–141. (b) Kim, D. Y.; Singh, N. J.; Lee, H. M.; Kim, K. S. *Chem.—Eur. J.* **2009**, *15*, 5598. (c) Kim, D. Y.; Lee, H. M.; Seo, J.; Shin, S. K.; Kim, K. S. *Phys. Chem. Chem. Phys.* **2010**, *12*, 5446.
- (5) U. S. Department of Energy Hydrogen Program Annual Merit Review Proceedings, Arlington, VA, May 18–22, 2009; http://www.hydrogen.energy.gov/annual_review09_proceedings.html.
- (6) (a) Rowsell, J. L. C.; Yaghi, O. M. *Angew. Chem.* **2005**, *117*, 4748–4758. (b) Rowsell, J. L. C.; Yaghi, O. M. *Angew. Chem., Int. Ed.* **2005**, *44*, 4670–4679. (c) Han, S. S.; Furukawa, H.; Yaghi, O. M.; Goddard, W. A., III *J. Am. Chem. Soc.* **2008**, *130*, 11580–11581.
- (7) Kubas, G. J. *Chem. Rev.* **2007**, *107*, 4152.
- (8) (a) Kim, Y. H.; Zhao, Y. F.; Williamson, A.; Heben, M. J.; Zhang, S. B. *Phys. Rev. Lett.* **2006**, *96*, 016102. (b) Zhao, Y.; Kim, Y.-H.; Dillan, A. C.; Heben, M. J.; Zhang, S. B. *Phys. Rev. Lett.* **2005**, *94*, 155504.
- (9) (a) Sun, Q.; Wang, Q.; Jena, P.; Kawazoe, Y. *J. Am. Chem. Soc.* **2005**, *127*, 14582. (b) Kiran, B.; Kandalam, A. K.; Jena, P. *J. Chem. Phys.* **2006**, *124*, 224703. (c) Zhou, W.; Yildirim, T.; Durgun, E.; Ciraci, S. *Phys. Rev. B* **2007**, *76*, 085434. (d) Weck, P. F.; Kumar, T. J. D.; Kim, E.; Balakrishnan, N. *J. Chem. Phys.* **2007**, *126*, 094703. (e) Kim, T. S.; Kim, K. J.; Jo, S. K.; Lee, J. *J. Phys. Chem. B* **2008**, *112*, 16431. (f) Bhattacharya, S.; Majumder, C.; Das, G. P. *J. Phys. Chem. C* **2008**, *112*, 17487. (g) Okamoto, Y. *J. Phys. Chem. C* **2008**, *112*, 17721. (h) Barman, S.; Sen, P.; Das, G. P. *J. Phys. Chem. C* **2008**, *112*, 19963. (i) Bhattacharya, S.; Majumder, C.; Das, G. P. *J. Phys. Chem. C* **2009**, *113*, 15783.
- (10) Hamaed, A.; Trudeau, M.; Antonelli, D. M. *J. Am. Chem. Soc.* **2008**, *130*, 6992.
- (11) Kim, Y.-H.; Sun, Y. Y.; Choi, W. I.; Kang, J.; Zhang, S. B. *Phys. Chem. Chem. Phys.* **2009**, *11*, 11400.
- (12) Park, M. H.; Lee, Y. S. *Chem. Phys. Lett.* **2010**, *488*, 7.
- (13) (a) Tarakeshwar, P.; Kumar, T. J. D.; Balakrishnan, N. *J. Phys. Chem.* **2008**, *112*, 2846. (b) Kumar, T. J. D.; Tarakeshwar, P.; Balakrishnan, N. *J. Chem. Phys.* **2008**, *128*, 194714. (c) Tarakeshwar, P.; Kumar, T. J. D.; Balakrishnan, N. *J. Chem. Phys.* **2009**, *130*, 114301.
- (14) (a) Kim, K. S.; Suh, S. B.; Kim, J. C.; Hong, B. H.; Lee, E. C.; Yun, S.; Tarakeshwar, P.; Lee, J. Y.; Kim, Y.; Ihm, H.; Kim, H. G.; Lee, J. W.; Kim, J. K.; Lee, H. M.; Kim, D.; Cui, C.; Youn, S. J.; Chung, H. Y.; Choi, H. S.; Lee, C.-W.; Cho, S. J.; Jeong, S.; Cho, J.-H. *J. Am. Chem. Soc.* **2002**, *124*, 14268. (b) Singh, N. J.; Lee, H. M.; Hwang, I.-C.; Kim, K. S. *Supramol. Chem.* **2007**, *19*, 321. (c) Singh, N. J.; Lee, E. C.; Choi, Y. C.; Lee, H. M.; Kim, K. S. *Bull. Chem. Soc. Jpn.* **2007**, *80*, 1437. (d) 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. (e) Chellappan, K.; Singh, N. J.; Hwang, I.-C.; Lee, J. W.; Kim, K. S. *Angew. Chem., Int. Ed.* **2005**, *44*, 2899.
- (15) Bushnell, J. E.; Kemper, P. R.; Maitre, P.; Bowers, M. T. *J. Am. Chem. Soc.* **1994**, *116*, 9710.
- (16) Bushnell, J. E.; Maitre, P.; Kemper, P. R.; Bowers, M. T. *J. Chem. Phys.* **1997**, *106*, 10153.
- (17) Bushnell, J. E.; Kemper, P. R.; Bowers, M. T. *J. Phys. Chem.* **1993**, *97*, 11628. (b) Maitre, P.; Bauschlicher, C. W., Jr. *J. Phys. Chem.* **1995**, *99*, 6836.
- (18) Bauschlicher, C. W., Jr.; Partridge, H.; Langhoff, S. R. *J. Phys. Chem.* **1992**, *96*, 2475.
- (19) Weis, P.; Kemper, P. R.; Bowers, M. T. *J. Phys. Chem.* **1997**, *101*, 2809.
- (20) Buchnell, J. E.; Kemper, P. R.; Bowers, M. T. *J. Phys. Chem.* **1995**, *99*, 15602.
- (21) (a) Kemper, P. R.; Bushnell, J.; von Helden, G.; Bowers, M. T. *J. Phys. Chem.* **1993**, *97*, 52. (b) Bauschlicher, C. W., Jr.; Maitre, P. *J. Phys. Chem.* **1995**, *99*, 3444.
- (22) (a) Kemper, P. R.; Weis, P.; Bowers, M. T.; Maitre, P. *J. Am. Chem. Soc.* **1998**, *120*, 13494. (b) Manard, M. J.; Bushnell, J. E.; Bernstein, S. L.; Bowers, M. T. *J. Phys. Chem. A* **2002**, *106*, 10027.
- (23) Liu, C. S.; Zeng, Z. *Phys. Rev. B* **2009**, *79*, 245419.
- (24) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Salvador, P.; Dannenberg, J. J.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Baboul, A. G.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 03*, revision A.1; Gaussian, Inc.: Pittsburgh, PA, 2003.
- (25) Lee, S. J.; Chung, H. Y.; Kim, K. S. *Bull. Korean Chem. Soc.* **2004**, *25*, 1061.
- (26) (a) Wachters, A. J. H. *J. Chem. Phys.* **1970**, *52*, 1033. (b) Bauschlicher, C. W., Jr.; Langhoff, S. R.; Barnes, L. A. *J. Chem. Phys.* **1989**, *91*, 2399.
- (27) Moore, C. E. *Ionization Potentials and Ionization Limits Derived from the Analyses of Optical Spectra*; National Bureau of Standards: Washington, DC, 1970.
- (28) Huber, K. P.; Herzberg, G. *Molecular Spectra and Molecular Structure. IV. Constants of Diatomic Molecules*; Van Nostrand Reinhold, Co.: New York, 1979, pp 99–116.
- (29) <http://www.chemicool.com/elements/titanium.html>.
- (30) (a) Helgaker, T.; Ruden, T. A.; Jorgensen, P.; Olsen, J.; Klopper, W. *J. Phys. Org. Chem.* **2004**, *17*, 913. (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.
- (31) Boese, A. D.; Oren, M.; Atasoylu, O.; Martin, L. M. L.; Kallay, M.; Gauss, J. *J. Chem. Phys.* **2004**, *120*, 4129.
- (32) Zhao, Y.; Lusk, M. T.; Dillon, A. C.; Heben, M. J.; Zhang, S. B. *Nano Lett.* **2008**, *8*, 157.
- (33) Wang, L.; Lee, K.; Sun, Y.-Y.; Lucking, M.; Chen, Z.; Zhao, J. J.; Zhang, S. B. *ACS Nano* **2009**, *3*, 3995.
- (34) Yildirim, T.; Ciraci, S. *Phys. Rev. Lett.* **2005**, *94*, 175501.
- (35) Durgun, E.; Ciraci, S.; Zhou, W.; Yildirim, T. *Phys. Rev. Lett.* **2006**, *97*, 226102.
- (36) Zhang, C.-G.; Zhang, R.; Wang, Z.-X.; Zhou, Z.; Zhang, S. B.; Chen, Z. *Chem.—Eur. J.* **2009**, *15*, 5910.

Electronic Structure and Effectively Unpaired Electron Density Topology in *closo*-Boranes: Nonclassical Three-Center Two-Electron Bonding

Rosana M. Lobayan,[†] Roberto C. Bochicchio,^{*,‡} Alicia Torre,[§] and Luis Lain[§]

[†]Facultad de Ingeniería, Universidad de la Cuenca del Plata, Lavalle 50, 3400, Corrientes, Argentina

[‡]Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, and Instituto de Física de Buenos Aires, Consejo Nacional de Investigaciones Científicas y Técnicas, Ciudad Universitaria, 1428, Buenos Aires, Argentina

[§]Departamento de Química Física, Facultad de Ciencia y Tecnología, Universidad del País Vasco, Apartado 644, E-48080 Bilbao, Spain

ABSTRACT: This article provides a detailed study of the structure and bonding in *closo*-borane cluster compounds $X_2B_3H_3$ ($X = BH^-, P, SiH, CH, N$), with particular emphasis on the description of the electron distribution using the topology of the quantum many-body effectively unpaired density. The close relationship observed between the critical points of this quantity and the localization of the electron cloud allows us to characterize the nonclassical bonding patterns of these systems. The obtained results confirm the suitability of the local rule to detect three-center two-electron bonds, which was conjectured in our previous study on boron hydrides.

1. INTRODUCTION

The physical information contained in N -electron molecular state functions is usually summarized by means of chemical descriptors which quantify the fundamental chemical concepts, such as atomic and bonding populations, bond orders and ionicities (bond multiplicities), atomic and free valencies, and others.^{1–12} The electron density expressed as electron population^{13–16} constitutes a tool of paramount importance in the treatments known as *nonlocal* or *integrated formulations*.^{4,13–16} Complementary to these methodologies are those based on the study of the electron density $\rho(\mathbf{r})$ and its associated Laplacian field $\nabla^2\rho(\mathbf{r})$ through their topological structures, i.e., the localization and classification of their critical points (cp's), (i.e., maxima, minima, or saddle ones). This approach to interpret chemical information from the local point of view^{17–19} will be called *local formulation*. A more detailed description of these methods of study may be found in ref 20.

An exact partitioning of the electron density has been reported so that $\rho(\mathbf{r}) = \rho^{(p)}(\mathbf{r}) + \rho^{(u)}(\mathbf{r})$ in which the contributions $\rho^{(p)}(\mathbf{r})$ and $\rho^{(u)}(\mathbf{r})$, possessing different physical characters, correspond to the effectively paired electron density and unpaired electron density, respectively.²¹ Both contributions to the electron density have been topologically described by the localization of their critical points, the spatial shifts of these cp's in comparison with those of the total density, and the determination of the domains in which these scalar fields concentrate/reduce.^{21,22}

It has been shown that the positions of the nuclear cp's (ncp's) (local maxima placed at the nuclear positions) of $\rho^{(p)}(\mathbf{r})$ are located very close to the total $\rho(\mathbf{r})$ ncp's, while those of $\rho^{(u)}(\mathbf{r})$ are also located close to the nuclear positions but out of the bonding region.²¹ Further studies dealing with the utilization of the Laplacian functions $\nabla^2\rho^{(p)}(\mathbf{r})$ and $\nabla^2\rho^{(u)}(\mathbf{r})$ stressed the fact that nuclear regions possess most of the effectively unpaired electrons. Moreover, both paired and unpaired density fields

possess successive regions of concentration and depletion of density, yielding a shell structure.²² These mentioned results have been obtained from the application of this local formalism to systems possessing conventional patterns of bonding.^{21,22} Recently, this treatment has successfully been extended to the description of the challenging electron distribution of systems with more complex bonding patterns, such as three-center two-electron (3c–2e) bonds in electron-deficient boron hydrides.²⁰ Following this line of research, our purpose is to continue studying the capability of our tools to properly describe complex structures of bonding. To this end we have chosen a class of evolving pattern boron structures,²³ the *closo*-borane family $X_2B_3H_3$ ($X = BH^-, P, SiH, CH, N$), for which classical and nonclassical alternatives of bonding have been suggested.^{24,25}

The organization of this article is as follows. Section 2 briefly reports the theoretical framework of the partitioning of the electron density, the relationships between the density gradients and between the Laplacian fields, and the tools used to carry out complementary studies of topological population analysis. Section 3 describes the computational details and the discussion of the calculations performed over the selected *closo*-borane $X_2B_3H_3$ structures. Section 4 is devoted to the concluding remarks.

2. THEORETICAL OUTLINE

2.1. Density Decomposition. The electron density, $\rho(\mathbf{r})$, in an N -electron molecular system may be decomposed into two contributions for both closed- and open-shell state functions:^{21,26}

$$\rho(\mathbf{r}) = \rho^{(p)}(\mathbf{r}) + \rho^{(u)}(\mathbf{r}) \quad (1)$$

Received: December 27, 2010

Published: March 10, 2011

where the effectively paired, $\rho^{(p)}(\mathbf{r})$, and unpaired, $\rho^{(u)}(\mathbf{r})$, densities are defined by

$$\rho^{(p)}(\mathbf{r}) = \frac{1}{2} \int d\mathbf{r}' {}^1D(\mathbf{r}|\mathbf{r}') {}^1D(\mathbf{r}'|\mathbf{r}) \quad (2)$$

and

$$\rho^{(u)}(\mathbf{r}) = \frac{1}{2} u(\mathbf{r}|\mathbf{r}) \quad (3)$$

respectively. ${}^1D(\mathbf{r}|\mathbf{r}')$ is the spin-free first-order reduced density matrix (1-RDM) in the coordinate representation; its trace is the number of electrons in the system, N .^{13,14} $u(\mathbf{r}|\mathbf{r})$ is the diagonal element of the effectively unpaired density matrix defined by

$$u(\mathbf{r}|\mathbf{r}') = 2{}^1D(\mathbf{r}|\mathbf{r}') - {}^1D(\mathbf{r}|\mathbf{r}) \quad (4)$$

where ${}^1D^2(\mathbf{r}|\mathbf{r}')$ is defined by the integral in eq 2.^{8–12} The densities are defined as the diagonal part of the corresponding density matrices.^{13,14} The traces of the effectively paired and unpaired densities are related to the number of paired (opposite spins) and unpaired electrons under the electron correlation effects.^{8,27} Two sources of unpaired density can be pointed out: one from the spin density (only present in nonsinglet states) and the other from the irreducible part or many-body cumulant of the second-order reduced density matrix (2-RDM)^{20,28} being supported by the Coulomb interaction between the particles.^{7,8,29}

Hence, for state functions having all orbitals doubly occupied, as in the closed-shell Hartree–Fock or the density functional theory (DFT) cases, $\rho^{(u)}(\mathbf{r})$ is intrinsically zero²¹ and cannot be considered for the electronic structure description of the systems.

2.2. Electron Density Topology: Local Information. The techniques used to study the electron density topology^{17,18} can be applied to the total density and its two contributions from the above-reported partition, in order to describe the bonding features in the local frame. The localization of the cp's of each density contribution, their values, and the corresponding increment/decrease of the densities in their surroundings provide enough information to describe the electron distribution.

The cp's of the total electron density are classified regarding two values associated with the Hessian matrix of $\rho(\mathbf{r})$: its *rank* r (number of nonzero eigenvalues) and its *signature* s (sum of signs of its three eigenvalues). This information is featured as (r,s) . A cp with all negative eigenvalues is denoted as $(3,-3)$ and is called the nuclear critical point (ncp); it indicates a local maximum and is placed very close to the nuclear positions. A cp with two negative eigenvalues and a positive one is denoted by $(3,-1)$ and stands for a bond critical point (bcp); this type of cp corresponds to a saddle point of the electron density and denotes a bonding interaction between two atoms.^{17,18} The first two eigenvalues of the Hessian matrix represent the curvatures of the densities along two axes perpendicular with respect to the third one associated with the internuclear axis. Within this framework, a covalent bond is featured by the existence of two large negative curvatures perpendicular to the bond line and a small positive curvature along the bond at the position of the bcp.^{17,18} Other important cp's are the *ring* (rcp) and *cage* (ccp) critical points, which usually appear in the complex electronic structure of molecular systems; these points are characterized by rank and signature $(3,+1)$ and $(3,+3)$, respectively.^{17,18} The value of $\nabla^2\rho(\mathbf{r})$ is the sum of the curvatures along the orthogonal coordinate axes; its sign indicates that the density is locally depleted (positive) or concentrated (negative)

Table 1. Local and Integrated (Nonlocal) Topological Features of the Total Density $\rho(\mathbf{r})$ for $X_2B_3H_3$ ($X = BH^-, P, SiH, CH, N$) *closo*-Borane Systems at the CISD/6-31G Level of Approximation^a**

cp type	ρ sequences ^b	bond	I_{Ω, Ω_b}	$\Delta_{\Omega, \Omega_b, \Omega_c}^{(3)}$
$X = BH^-$				
ncp	one on each B, B', H atom			
bcp	one per each of BH bond	BB	0.404	
	one per each of B'H bond	BB'	0.786	
	one per each of BB' bond	BH	0.068	
	no BB bonds present	B'H	0.681	
	no B'B' bonds present	B'B'	0.135	
rcp	one per each BB'BB' sequence	BBB		0.043
ccp	one BB'BB'B	BB'B		0.201
		B'BB'		0.000
$X = P$				
ncp	one on each P, B, H atom			
bcp	one per each of BP bond	BP	0.853	
	one per each of BH bond	BH	0.666	
	no BB bonds present	BB	0.257	
	no PP bonds present	PP	0.317	
rcp	one per each BPBP	BBB		0.018
ccp	one PBBBB	BPB		0.177
		PBP		0.046
$X = SiH$				
ncp	one on each B, Si, H atom			
bcp	one per each of BSi bond	BSi	0.591	
	one per each of BH bond	BH	0.782	
	one per each of SiH bond	SiH	0.590	
	no BB bonds present	BB	0.452	
	no SiSi bonds present	SiSi	0.036	
rcp	one per each BSiBSi sequence	BBB		0.060
ccp	one SiBBBBi	BSiB		0.121
		SiBSi		0.022
$X = CH$				
ncp	one on each B, C, H atom			
bcp	one per each of BC bond	BC	0.485	
	one per each of BH bond	BH	0.547	
	one per each of CH bond	CH	1.003	
	no BB bonds present	BB	0.026	
	no CC bonds present	CC	0.697	
rcp	one for each BCBCB sequence	BBB		0.001
ccp	one CBBBC	BCB		0.031
		CBC		0.000
$X = N$				
ncp	one on each N, B, H atom			
bcp	one per each of BN bond	BN	0.411	
	one per each of BH bond	BH	0.527	
	no BB bonds present	BB	0.017	
	no NN bonds present	NN	0.786	
rcp	one per each BNB sequence	BBB		0.001
ccp	one NBBBN	BNB		0.021
		NBN		0.081

^a All quantities are in atomic units. ^b Indicates the nucleus at which the ncp is located; for bcp's, the atoms defining the bond; for rcp's, the atoms giving rise to the ring; for ccp's, the atoms defining the cage. ^c Equatorial and vertex boron atoms are symbolized by B and B', respectively.

and constitutes valuable information to describe the behavior of the density around a point.^{17–19} According to critical point definition, the total density cp's are found by means of the

Table 2. Local and Integrated (Nonlocal) Topological Features of the Effectively Unpaired Density $\rho^{(u)}(\mathbf{r})$ for $X_2B_3H_3$ ($X = BH^-, P, SiH, CH, N$) *closo*-Borane Systems at the CISD/6-31G Level of Approximation^a**

cp type	$\rho^{(u)}$ sequences ^b	atom	$\mu_{\Omega_\lambda}^c$
$X = BH^-^d$			
vs (3,-3) cp	one on each B, B', H, H' atom	B	0.154
		B'	0.162
		H	0.079
		H'	0.080
vs (3,-1) cp	one per each BB' sequence one per each BH and BH' sequence one per each BB sequence		
vs (3,+1) cp	one per each BB'B sequence		
vs (3,+3) cp	one B'BBBB'		
$X = P$			
vs (3,-3) cp	one on each B, H atom one for each P atom	B	0.135
		P	0.304
		H	0.073
vs (3,-1) cp	one per each BP sequence one per each BH sequence one per each BB sequence		
vs (3,+1) cp	one for each BPB sequence one BBB		
vs (3,+3) cp	one for each PBBB sequence		
$X = SiH$			
vs (3,-3) cp	one on each B, H atom one for each Si atom	B	0.200
		Si	0.108
		H (BH)	0.075
		H (SiH)	0.076
vs (3,-1) cp	one per each BH' sequence one per each BH sequence one per each BB sequence		
vs (3,+1) cp	one for each BB sequence two for each BBB sequence one per each BH'B sequence		
vs (3,+3) cp	one SiBBBBi		
$X = CH$			
vs (3,-3) cp	one on each B, H atom one for each C atom	B	0.063
		C	0.302
		H (BH)	0.069
		H (CH)	0.044
vs (3,-1) cp	one per each BC sequence one per each BH sequence one per each BB sequence		
vs (3,+1) cp	one for each BCB sequence		
vs (3,+3) cp	one CBBBC		
$X = N$			
vs (3,-3) cp	one on each B, H atom one for each N atom	B	0.053
		N	0.342
		H	0.066
vs (3,-1) cp	one per each BN sequence one per each BH sequence one per each BB sequence		
vs (3,+1) cp	one for each BNB sequence		
vs (3,+3) cp	one NBBBN		

^a All quantities are in atomic units. ^b Indicates the nucleus at which the vs (3,-3) cp is located; for vs (3,-1) cp's, the atoms defining the bond; for vs (3,+1) cp's, the atoms giving rise to the ring; for vs (3,+3) cp's, the atoms defining the cage. ^c Effectively unpaired atomic electron index. ^d Equatorial and vertex borons and their linked hydrogen atoms are symbolized by B, B' and H, H', respectively.

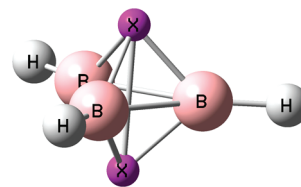


Figure 1. Geometric conformation of *closo*-borane systems. Atomic symbols are shown; $X = BH^-, P, SiH, CH, N$.

Table 3. Density and $L(r)$ of the Total Density ρ and $\rho^{(u)}$ for $X_2B_3H_3$ ($X = BH^-, P, SiH, CH, N$) *closo*-Borane Systems at Bond Critical Points of the Total Density at CISD/6-31G Level of Calculation^{a,b}**

system ^c	bond	$\rho(r) _{bcp}$	$\rho^{(u)}(r) _{bcp}$	$-\nabla^2\rho(r) _{bcp}$	$-\nabla^2\rho^{(u)}(r) _{bcp}$	ϵ^d	
$(B'H^-)_2(BH)_3$	BB'	0.1354	0.0033	0.1913	0.0010	0.4210	
		0.1349	0.0034	0.2028	0.0012	0.0837	
	BB'	0.0962	0.0024	-0.0030	-0.0014	0.5539	
	BH	0.1482	0.0033	-0.0294	0.0098	0.3686	
		0.1650	0.0032	0.4349	0.0043	0.1832	
	B'H	0.1470	0.0033	0.0584	0.0085	0.0000	
$P_2(BH)_3$	BP	0.1640	0.0031	0.4278	0.0040	0.0000	
	BP	0.1303	0.0032	0.1947	0.0125	0.1830	
	BB'	0.1308	0.0038	0.1942	0.0011	0.1087	
	BB'	0.0847	0.0023	-0.0130	-0.0008	0.2128	
BH	BH	0.1807	0.0037	0.2091	0.0117	0.2186	
		0.1455	0.0042	-2.1590	0.1358	8.3965	
	$(SiH)_2(BH)_3$	BSi ^g	0.1006	0.0023	-0.0488	0.0063	0.3336
	BB'	0.0773	0.0021	-0.1705	0.0076	8.1221	
BH	BH	0.1170	0.0019	-0.2584	0.0062	0.0000	
		0.1966	0.0035	0.6285	0.0038	0.1449	
	BH' ^f	0.0650	0.0016	0.2483	-0.0054	0.1307	
	SiH' ^g	0.1170	0.0019	-0.2584	-0.0054	0.1307	
$(CH)_2(BH)_3$	BC	0.1676	0.0042	-0.0582	0.0127	0.0213	
		0.1845	0.0039	0.3841	0.0014	0.0397	
	BB'	0.0972	0.0230	-0.0987	-0.0021	2.0961	
	BH	0.1818	0.0035	0.2050	0.0115	0.2014	
CH	CH	0.1970	0.0032	0.6330	0.0032	0.1389	
	CH	0.2893	0.0046	1.0788	-0.0039	0.0000	
		0.2900	0.0045	1.1413	-0.0033	0.0000	
	$N_2(BH)_3$	BN	0.1924	0.0044	-0.5671	0.0144	0.0490
BH	BH	0.2155	0.0044	0.3931	0.0047	0.0581	
	BB'	0.1172	0.0029	-0.1748	-0.0008	1.9502	
	BH	0.1894	0.0034	0.2745	0.0121	0.1229	
		0.2043	0.0032	0.6853	0.0026	0.1012	

^a Second row in columns 3-7 for each bond indicates the densities and $L(r)$ at $\rho^{(u)}(r)$ vs (3,-1) cp. ^b All quantities are in atomic units. ^c See Figure 1 for atom labeling. ^d Ellipticity. ^e B', B, vertex and equatorial boron atoms, respectively. ^f There are no bcp points for $\rho(r)$ between these atoms. ^g There are no vs (3,-1) cp's for $\rho^{(u)}(r)$ between these atoms.

gradient of the field by

$$\nabla\rho(r)|_{r^c} = 0 \quad (5)$$

and according to eq 1

$$\nabla\rho^{(p)}(r)|_{r^c} + \nabla\rho^{(u)}(r)|_{r^c} = 0 \quad (6)$$

where $r^c = \{r_i^c; i = 1, \dots, M\}$ indicates the set of critical points of

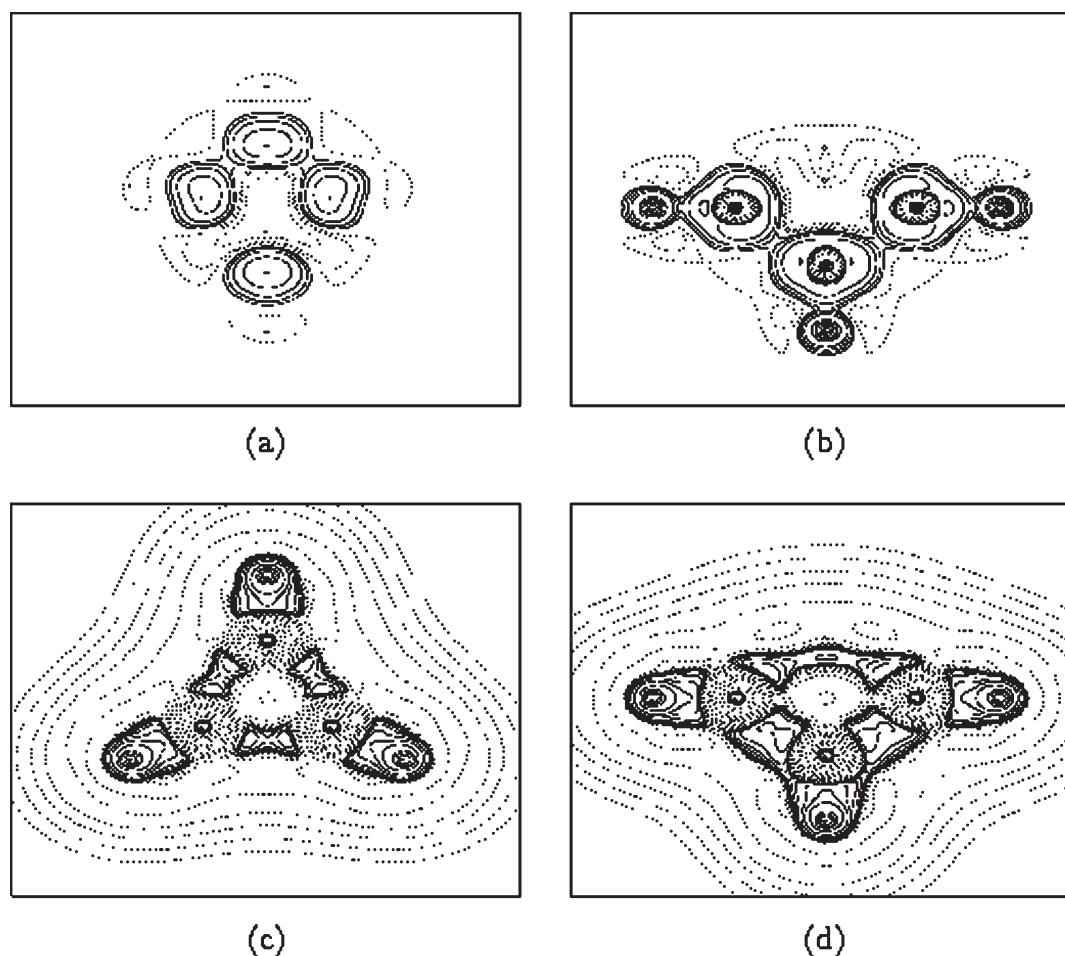


Figure 2. $L(\mathbf{r})$ contour maps of effectively unpaired (a, b) and paired (c, d) densities of the $(\text{BH}^-)_2(\text{BH})_3$ system: (a) on the plane containing two equatorial boron atoms (B_{eq}) and an $\text{X} = \text{B}$ atom; (b) on the plane of the $\text{BB}_{\text{eq}}\text{B}$ atoms; (c) on the plane containing the three equatorial boron atoms (B_{eq}); (d) on the plane of the $\text{BB}_{\text{eq}}\text{B}$ atoms. Positive and negative values are denoted by solid and dashed lines, respectively.

the total density $\rho(\mathbf{r})$. As is obvious, the following relation holds:

$$\nabla \rho^{(\text{p})}(\mathbf{r})|_{\text{rc}} = -\nabla \rho^{(\text{u})}(\mathbf{r})|_{\text{rc}} \quad (7)$$

whose physical meaning is that each density field component increases/decreases its value in opposite direction. Hence, no simultaneous increment/decrease of each one may occur at the cp's.²¹ Nevertheless, the Laplacian field of eq 1 yields

$$\nabla^2 \rho(\mathbf{r})|_{\text{rc}} = \nabla^2 \rho^{(\text{p})}(\mathbf{r})|_{\text{rc}} + \nabla^2 \rho^{(\text{u})}(\mathbf{r})|_{\text{rc}} \neq 0 \quad (8)$$

indicating that both $\nabla^2 \rho^{(\text{p})}(\mathbf{r})|_{\text{rc}}$ and $\nabla^2 \rho^{(\text{u})}(\mathbf{r})|_{\text{rc}}$ contributions do not necessarily follow opposite trends. Hence, both densities may simultaneously concentrate or deplete in the neighborhood of a cp.²⁰ The terminology vs (3,−1) cp, vs (3,+1) cp, and vs (3,+3) cp will refer to (3,−1), (3,+1), and (3,+3) critical points of the $\rho^{(\text{u})}(\mathbf{r})$ valence shell, in analogy with the bcp, rcp, and ccp's of the total density, respectively. Nevertheless, it is important to note that such points are not *sensu strictu* bcp, rcp, or ccp's because only the cp's of the total density are able to define a bond in the AIM topological formalism.^{17,18}

2.3. Electron Density Topology: Nonlocal Information.

The nonlocal or integrated formalism is complementary to the above-mentioned local one. This nonlocal formalism relates the classical chemical concepts such as atomic charges, bond orders and valences, etc. to physical magnitudes that quantify them, i.e.,

chemical indicators.^{7,8,29–31} The relations defining the relevant magnitudes to our goal within the nonlocal AIM topological population analysis are the two-center bond index defined as

$$I_{\Omega_A \Omega_B} = \sum_{i,j,k,l} {}^1D_j^i {}^1D_l^k S_{il}(\Omega_A) S_{kj}(\Omega_B) \quad (9)$$

where Ω_A and Ω_B stand for Bader's atomic domains in the physical space,¹⁷ ${}^1D_j^i$ are the first-order reduced density matrix elements, and $S_{ij}(\Omega_A)$ are the elements of the overlap matrix over the region Ω_A in the orthogonal molecular basis set $\{i, j, k, l, \dots\}$,^{7,29} and the three-center bond index as³⁰

$$I_{\Omega_A \Omega_B \Omega_C} = \sum_{i,j,k,l,m,n} {}^1D_j^i {}^1D_l^k {}^1D_n^m S_{in}(\Omega_A) S_{kj}(\Omega_B) S_{ml}(\Omega_C) \quad (10)$$

expressing the three-center topological bonding populations by

$$\Delta_{\Omega_A \Omega_B \Omega_C}^{(3)} = \frac{1}{4} \sum_{P(\Omega_A \Omega_B \Omega_C)} I_{\Omega_A \Omega_B \Omega_C} \quad (11)$$

where $P(\Omega_A \Omega_B \Omega_C)$ indicates the permutations of the three domain contributions.

As pointed out in the Introduction, the main purpose of this paper is to study the capability of these tools to describe

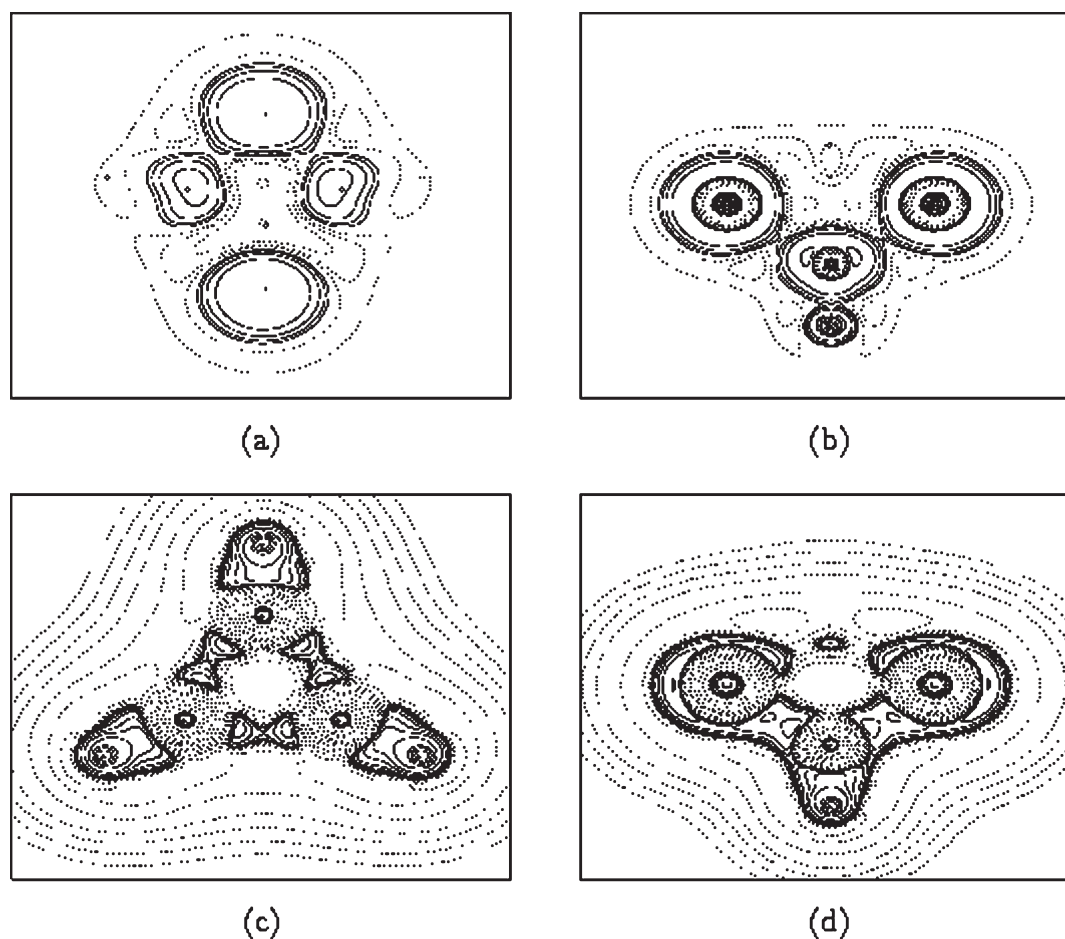


Figure 3. $L(\mathbf{r})$ contour maps of effectively unpaired (a, b) and paired (c, d) densities of the $P_2(BH)_3$ system: (a) on the plane containing two equatorial boron atoms (B_{eq}) and an $X = P$ atom; (b) on the plane of the $PB_{eq}P$ atoms; (c) on the plane containing the three equatorial boron atoms (B_{eq}); (d) on the plane of the $PB_{eq}P$ atoms. Positive and negative values are denoted by solid and dashed lines, respectively.

nonclassical patterns of bonding such as the $3c-2e$, among others. The results are reported in section 3.

3. COMPUTATIONAL DETAILS, RESULTS, AND DISCUSSION

As is well-known, the boron atom forms electron-deficient compounds giving rise to $3c-2e$ bonds in boron hydrides.²³ A previous study permitted noting that the onset of such a bonding character seems to be closely related to the effectively unpaired electron density distribution featured by the unpaired electron delocalization on the bonding regions, its accumulation on the bcp's, and the existence of vs $(3, -1)$ cp's and vs $(3, +1)$ cp's of $\rho^{(u)}(\mathbf{r})$ between the atoms involved.²⁶ Therefore, the next step is to search for applications that would reveal the physical meaning of these characteristics in more complex structures. To this end, we have chosen a family of boron compounds, i.e., the *closo*-borane clusters $X_2B_3H_3$ ($X = N, CH, P, BH^-, SiH$),^{23,24} which incorporate a new feature, i.e., additional electrons provided by the X group. These types of systems are natural candidates to possess $3c-2e$ patterns of bonding. Hence, this scenario is adequate to enlighten the nature of this pattern passing from typical boron hydrides to more complex borane systems which seem to have nonclassical bonding structures.²⁴

The state functions used in this work to describe the selected molecular systems in their singlet ground states were calculated at the level of configuration interaction with single and double excitations (CISD) and the singlet RHF state as reference, using the Gaussian 03 package³² with the basis sets 6-31G**. The geometries for all systems were optimized within this approximation. The densities, their critical points, and their Laplacian fields $\nabla^2\rho(\mathbf{r})$ and $\nabla^2\rho^{(u)}(\mathbf{r})$ were determined by appropriately modified AIMPACK modules.³³ The numerical results of the electron population analysis were obtained with our own codes.³⁰ For practical reasons, we will use the function $L(\mathbf{r}) = -\nabla^2\rho(\mathbf{r})$ in the discussion of results as an indicator of local concentration (positive value) or local depletion (negative value) of the number of electrons at the point \mathbf{r} ;^{20-22,26} the terms "accumulation" and "reduction" have been proposed for the description of maxima and minima in $\rho(\mathbf{r})$.^{18,35} Because of the complex structure of the $\rho^{(u)}(\mathbf{r})$ topology, we will only deal in our study with critical points associated with its valence shells (vs) in the corresponding systems and no reference will be made to those of the inner shells of this density; in fact, only the former ones are involved in bonding phenomena.

Before discussing the results, let us enunciate the previously established quantum based rule to detect $3c-2e$ bonds in borane clusters to be applied to the complex structure of *closo*-borane

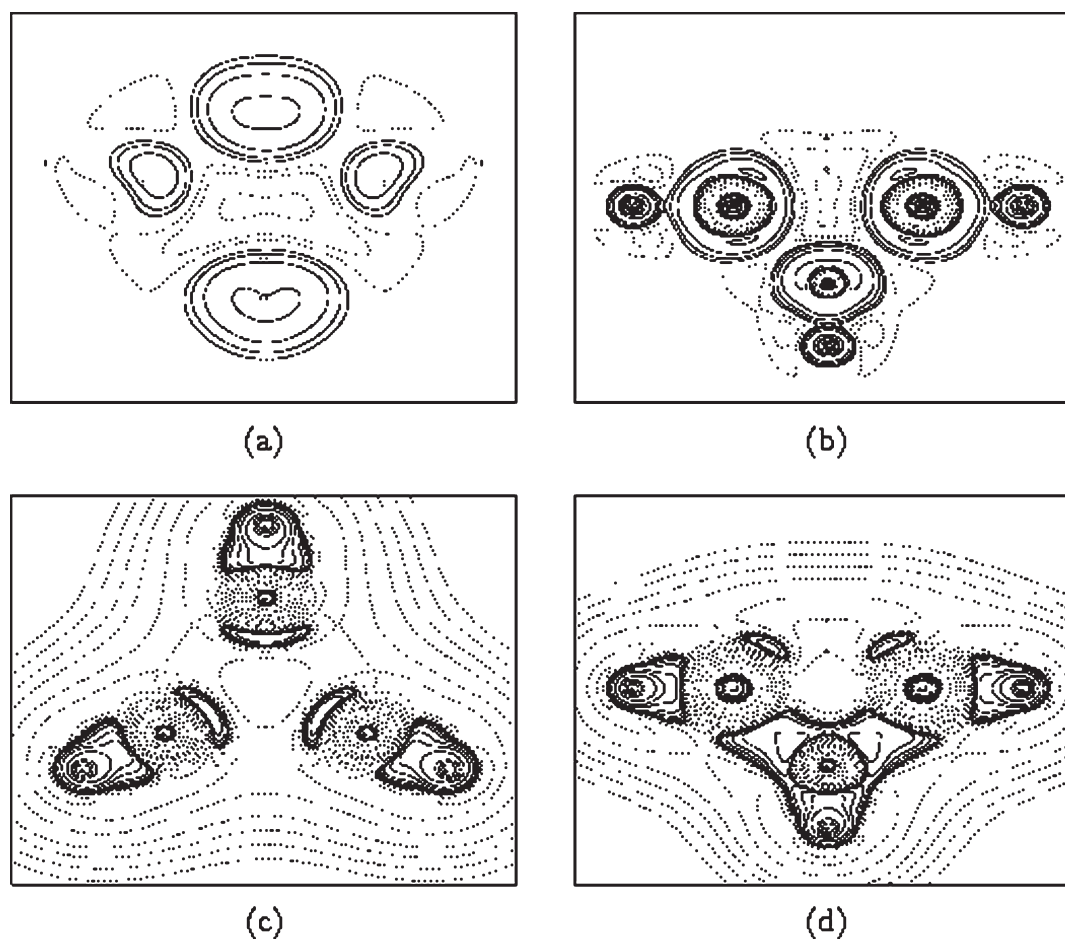


Figure 4. $L(r)$ contour maps of effectively unpaired (a, b) and paired (c, d) densities of the $(\text{SiH})_2(\text{BH})_3$ system: (a) on the plane containing two equatorial boron atoms (B_{eq}) and an $X = \text{Si}$ atom; (b) on the plane of the $\text{SiB}_{\text{eq}}\text{Si}$ atoms; (c) on the plane containing the three equatorial boron atoms (B_{eq}); (d) on the plane of the $\text{SiB}_{\text{eq}}\text{Si}$ atoms. Positive and negative values are denoted by solid and dashed lines, respectively.

molecular systems, in a general manner as follows: a $3c-2e$ bond between atoms ABC exists if there is a $vs(3,-1)$ cp of $\rho^{(u)}$ between each pair of atoms involved in the three-center ABC sequence and a $vs(3,+1)$ cp defined only by the atoms involved in the three-center bond, hereafter called the *local rule*.²⁰ This result seems to be the local version of the criterion of the integrated formalism of population analysis for detecting three-center bonding, hereafter the *integrated* or *nonlocal rule*. That criterion settles the existence of a $3c-2e$ bond between atoms ABC when fractional bond orders $I_{\Omega_A\Omega_B}$ appear between all possible pairs of atoms AB , BC , and AC and an appreciable $\Delta_{\Omega_A\Omega_B\Omega_C}^{(3)}$ defines its strength.^{4,36} Because of the detailed description of the distribution from the topological (local) view, we will adopt this form of the rule as the indicator of the quality of the interaction between the atoms while the integrated form is interpreted as an indicator of the strength of such interaction; i.e., the local rule defines the existence of such a type of interaction between the atoms while the integrated (nonlocal) form of the rule adjudicates the strength of the interaction by means of the population shared by the atoms involved.

Figure 1 shows the geometric conformation of the *closo*-borane cluster compounds $\text{X}_2\text{B}_3\text{H}_3$ ($X = \text{N}, \text{CH}, \text{P}, \text{BH}^-, \text{SiH}$).^{23-25,34} Three boron atoms are located at the equatorial plane, each one bonded to a hydrogen atom. The X vertex moieties are symmetrically placed above and below the

equatorial plane.²³ Table 1 contains the topological information concerning the total density $\rho(r)$, showing the localization of the cp 's, their type and the atomic sequence which defines each of them, the two-center bond indices $I_{\Omega_A\Omega_B}$, and the three-center bonding electron populations $\Delta_{\Omega_A\Omega_B\Omega_C}^{(3)}$. Table 2 is devoted to the topological structure of $\rho^{(u)}(r)$ and the number of effectively unpaired electron u_{Ω_A} values. Note that no reference to the effectively paired density $\rho^{(p)}(r)$ is made in Table 2 because as shown in previous articles, its structure is similar to that of $\rho(r)$ and therefore it does not introduce any new information.^{21,22}

As mentioned above, Table 1 summarizes the main parameters describing the structure of the systems. It reports the nonnegligible values of the two-center covalent bond indices $I_{\Omega_A\Omega_B}$ and the three-center populations $\Delta_{\Omega_A\Omega_B\Omega_C}^{(3)}$. The information contained in Table 1 indicates that the boron atoms in the equatorial plane of the molecules are not bonded to each other, i.e., there is no bcp defined by two of them in any of these systems. However, they are bonded to the heavy atom in the X moieties. Nonnegligible two-center indices are obtained between the heavy atom of both X groups for $\text{X}_2\text{B}_3\text{H}_3$, $X = \text{CH}, \text{N}$ (approximately 0.7 and 0.8, respectively); however, no bonding character appears because there is no bcp defined between them. This means that the electron density accumulates in a spatial region between these atoms but it does not constitute a bond. Such an amazing feature will be discussed in more detail later. An

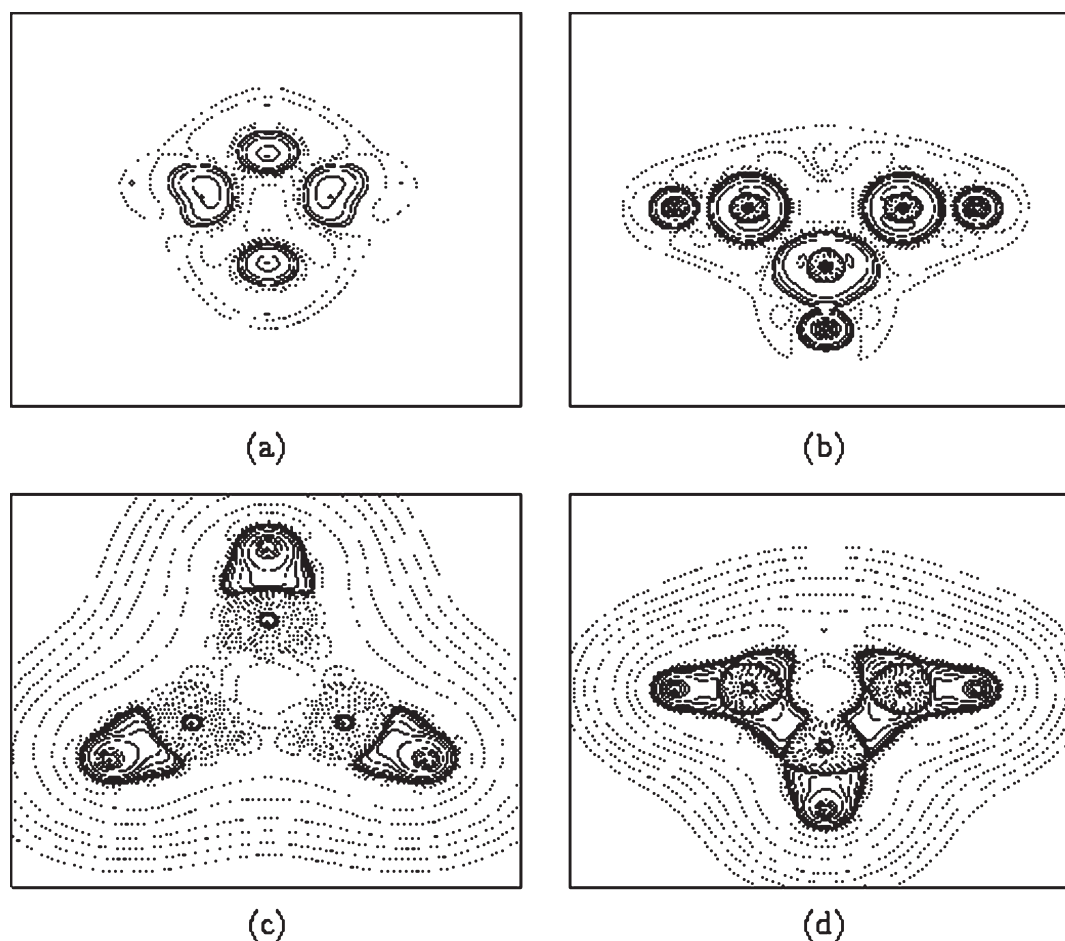


Figure 5. $L(r)$ contour maps of effectively unpaired (a, b) and paired (c, d) densities of the $(\text{CH})_2(\text{BH})_3$ system: (a) on the plane containing two equatorial boron atoms (B_{eq}) and an $\text{X} = \text{C}$ atom; (b) on the plane of the $\text{CB}_{\text{eq}}\text{C}$ atoms; (c) on the plane containing the three equatorial boron atoms (B_{eq}); (d) on the plane of the $\text{CB}_{\text{eq}}\text{C}$ atoms. Positive and negative values are denoted by solid and dashed lines, respectively.

rcp for each sequence formed by two of the equatorial boron atoms and the heavy atom of both X moieties is present in the systems. All the systems possess one ccp formed by these two atoms and the three equatorial boron atoms. Regarding the three-center populations, we may note that the BBB equatorial sequence shows low values indicating a small shared population of 0.043, 0.018, and 0.060 for $\text{X} = \text{BH}^-$, P, and SiH, respectively, and a negligible one for $\text{X} = \text{CH}$ and N. However, the sequence BXB shows such a type of bonding population increasing from 0.021 to 0.201 according to $\text{N} < \text{CH} < \text{SiH} < \text{P} < \text{BH}^-$. Some of the systems also show XBX nonnegligible populations, indicating weak $3\text{c}-2\text{e}$ bonds for $\text{X} = \text{P}$, SiH, and N. The cases of $\text{X} = \text{BH}^-$ and CH do not show this bonding character.

Table 2 collects the information from $\rho^{(u)}(\mathbf{r})$ required to apply the local form of the rule to detect $3\text{c}-2\text{e}$ bonds.²⁰ Hence, following this rule, all systems except $\text{X} = \text{SiH}$ possess $3\text{c}-2\text{e}$ bonds formed by the heavy atom of the X moiety and two equatorial boron atoms, i.e., BXB; the strength of each of these bonds is stated by the corresponding three-center population collected in Table 1, as reported above. The unpaired electron distribution around the Si atom is particularly complex, and it is not possible to assign a vs (3, -1) cp to the BX interactions in this case. However, the appearance of a vs (3, -1) cp for the sequences BB and BH' (where H' represents the hydrogen atom in the $\text{X} = \text{SiH}$ moiety) and a vs (3, +1) cp for each $\text{BH}'\text{B}$

sequence may be noted. These last features could explain the 0.121 three-center population reported in Table 1 for the BSiB sequence of this complex system. The observation of the increment of the u_{Ω} populations of the boron atoms in the equatorial plane following the sequence $\text{N} < \text{CH} < \text{P} < \text{BH}^- < \text{SiH}$ and a corresponding decrease of that of X's (see Table 2) may be interpreted as a transference of unpaired electron population from the three equatorial boron atoms toward the heavy atoms in the X moieties in all systems considered.

Table 3 and Figures 2–6 permit completion of the description of the above featured behavior of the electron distribution in the reported systems. Table 3 collects the values of the total and the unpaired densities and their associated Laplacian fields at the neighborhoods of the bcp's and vs (3, -1) cp's. Figures 2a,b–6a, b show the unpaired density maps on the planes defined by two equatorial boron atoms and the heavy atom of the X moiety and on the plane defined by the heavy atom in both X moieties and an equatorial boron atom, respectively. The paired density maps are shown on the equatorial plane in Figures 2c–6c, and Figures 2d–6d show this density at the same plane as Figures 2b–6b. The sign of the $L(r)$ function in Table 3 reveals concentration of both total and unpaired densities at the BX bcp and the vs (3, -1) cp for the $\text{X} = \text{BH}^-$, P systems; Figures 2a,b and 3a,b show that such concentration of $\rho^{(u)}(\mathbf{r})$ spills on the inter-nuclear bonding spatial regions of these systems, as previously

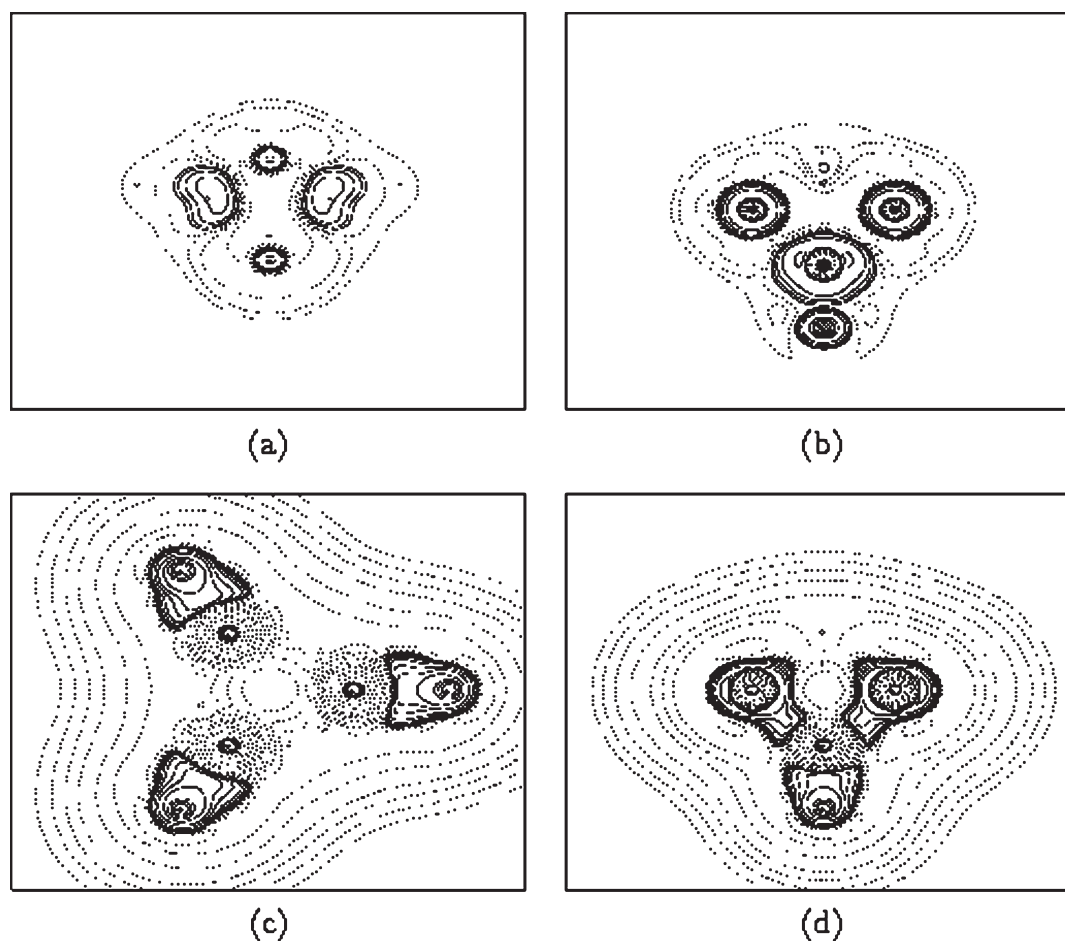


Figure 6. $L(r)$ contour maps of effectively unpaired (a, b) and paired (c, d) densities of the $N_2(BH)_3$ system: (a) on the plane containing two equatorial boron atoms (B_{eq}) and an $X = N$ atom; (b) on the plane of the $NB_{eq}N$ atoms; (c) on the plane containing the three equatorial boron atoms (B_{eq}); (d) on the plane of the $NB_{eq}N$ atoms. Positive and negative values are denoted by solid and dashed lines, respectively.

noted for boron hydrides.²⁰ The remaining systems, $X = SiH$, CH , and N , show marked depletion of $\rho(r)$ at the BX bcp, while $\rho^{(u)}(r)$ remains concentrated at the $vs(3, -1)$ cp's (cf. Table 3); i.e., there is an important shift of the bcp and $vs(3, -1)$ cp; Figures 4a,b, 5a,b, and 6a,b complement this result showing the unpaired density depletion in the spatial region between the above-mentioned atoms and its accumulation close to the nuclear positions. Hence, $\rho^{(u)}(r)$ becomes accumulated in all systems at bcp's, while $\rho(r)$ does not. This behavior may not show either ionic or covalent trends,²² which indicates a more complex type of bonding as described previously for boron hydrides.²⁰

The results described above evidence the presence of complex patterns of bonding, i.e., the existence of $3c-2e$ bonds defined by the BXB atoms. However, there is no evidence of $3c-2e$ BBB bonds. These reasons lead to consideration of nonclassical patterns of bonding for the electronic structure of all systems treated in this work (cf. ref 24). An important feature within these nonclassical structures that deserves discussion to complete the understanding of the electronic structure of these systems is related to the high value of some covalent two-center bond indices, i.e., between two equatorial boron atoms or between two heavy atoms, one on each X moiety, despite no bcp defining a true bond between them. This may be considered as a feature of the electron distribution that relates to some extent the nonlocal

and the local formulations of description. Let us begin with the first of the mentioned sequences, that is, the BB equatorial boron atoms. The $I_{\Omega_B\Omega_B}$ value increases following the $N < CH < P < BH^- < SiH$ ordering. To explain this behavior, we may note the contour maps of the $L(r)$ function of the effectively paired densities on the equatorial plane of the molecule (Figures 2c–6c), bearing in mind that integration of this density inside the corresponding atomic basins determines the two-center index (cf. eq 9). The systems with $X = SiH$, BH^- , and P moieties have appreciable values of 0.452, 0.404, and 0.257, respectively, for these $I_{\Omega_B\Omega_B}$ indices, in agreement with the pair density accumulation in the internuclear regions shown in Figures 2c–6c. Very small bond indices of 0.026 and 0.017 for $X = CH$ and N , respectively, are due to the marked depletion of the density in the mentioned regions. A similar analysis may be performed for the bondings between the heavy atoms of the X moieties, graphically shown in Figures 2d–6d. Finally, it is worth noting the relation between the covalent bond indices $I_{\Omega_B\Omega_B}$ and $I_{\Omega_X\Omega_X}$. It may be noted that $I_{\Omega_B\Omega_B}$ increases from 0.017 to 0.452 following the sequence mentioned above, while $I_{\Omega_X\Omega_X}$ decreases following the inverse sequence from 0.786 to 0.036. Therefore, the equatorial boron atom bonding populations decrease according to the increase of the bonding electron population of the heavy atom in the X moiety, in agreement with the increase of its electronegativity.

4. CONCLUDING REMARKS

It is well-known that B atoms tend to form electron-deficient compounds, particularly in boron hydrides. In this work, we have dealt with *closo*-borane molecular systems, where additional electrons are available and therefore such electron deficiency is broken. Our main goal has been to characterize and to test our previously stated rule to detect and describe systems possessing $3c-2e$ complex patterns of bonding. From the present study we may conclude that the topological description of the local rule has been successfully applied to these type of systems and its nonlocal counterpart acts as an indicator of the strength of this interaction. Although the systems are not electron deficient, $\rho^{(u)}(\mathbf{r})$ remains spilled on the spatial bonding regions as noted previously in boron hydrides. Besides, it may be pointed out that the only knowledge of the total density does not provide a complete description of the electron structure of these type of systems. Therefore, the topology of $\rho^{(u)}(\mathbf{r})$ supplies some new type of information regarding the structure of the electron distributions of complex systems.

A result dealing with the strength of the covalent bond order $I_{\Omega_A\Omega_B}$ and the existence of a bond critical point bcp which defines the linkage between atoms in the molecule may be remarked: an appreciable $I_{\Omega_A\Omega_B}$ does not always imply the existence of a true bond (for instance $I_{\Omega_N\Omega_N} = 0.786$ in $N_2B_3H_3$ but there is no bcp between two atoms). This is an amazing result that merits a further detailed study on systems with complex patterns of bonding. Extensions of this type of analysis which exploits the information contained in $\rho^{(u)}(\mathbf{r})$ to understand other complex bonding systems such as organometallic and hydrogen bonded compounds are being considered in our laboratories.

AUTHOR INFORMATION

Corresponding Author

*E-mail: rboc@df.uba.ar.

ACKNOWLEDGMENT

This report has been financially supported by Projects X017 (Universidad de Buenos Aires), PIP No. 11220090100061 (Consejo Nacional de Investigaciones Científicas y Técnicas, República Argentina), the Spanish Ministry of Education (Grant CTQ2009-07459/BQU), and the Universidad del País Vasco (Grant GIU09/43). We thank the Universidad del País Vasco for allocation of computational resources. R.M.L. acknowledges aid from Universidad de la Cuenca del Plata (Corrientes, Argentina) for facilities during the course of this work.

REFERENCES

- (1) Giambiagi, M.; Giambiagi, M. S.; Grepel, D. R.; Heymann, C. D. *J. Chim. Phys.* **1975**, *72*, 15.
- (2) Mayer, I. *Chem. Phys. Lett.* **1983**, *97*, 270.
- (3) Mayer, I. *Int. J. Quantum Chem.* **1986**, *29*, 73.
- (4) Boichichio, R. C. *THEOCHEM* **1991**, *228*, 209 and references therein.
- (5) Boichichio, R. C.; Lain, L.; Torre, A. *Chem. Phys. Lett.* **2003**, *374*, 567 and references therein.
- (6) Alcoba, D. R.; Boichichio, R. C.; Lain, L.; Torre, A. *Chem. Phys. Lett.* **2007**, *442*, 157.
- (7) Boichichio, R. C.; Lain, L.; Torre, A. *Chem. Phys. Lett.* **2003**, *375*, 45.
- (8) Boichichio, R. C. *THEOCHEM* **1998**, *429*, 229.
- (9) Lain, L.; Torre, A.; Boichichio, R. C.; Ponec, R. *Chem. Phys. Lett.* **2001**, *346*, 283.

- (10) Takatsuka, K.; Fueno, T.; Yamaguchi, K. *Theor. Chim. Acta* **1978**, *48*, 175.
- (11) Takatsuka, K.; Fueno, T. *J. Chem. Phys.* **1978**, *69*, 661.
- (12) Staroverov, V. N.; Davidson, E. R. *Chem. Phys. Lett.* **2000**, *330*, 161.
- (13) McWeeny, R. *Methods of Molecular Quantum Mechanics*; Academic: London, 1969; pp 115–158 and references therein.
- (14) Davidson, E. R. *Reduced Density Matrices in Quantum Chemistry*; Academic: New York, 1976; pp 57–96 and references therein.
- (15) Bamzai, A. S.; Deb, B. M. *Rev. Mod. Phys.* **1981**, *53*, 95.
- (16) Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure*; Macmillan: New York, 1982; pp 203–205 and references therein.
- (17) Bader, R. F. W. *Atoms in Molecules: A Quantum Theory*; Clarendon Press: Oxford, U.K., 1994; pp 13–52 and references therein.
- (18) Popelier, P. L. A. *Atoms in Molecules: An Introduction*; Pearson: London, 1999; pp 70–80.
- (19) Bader, R. F. W. *Chem.—Eur. J.* **2006**, *12*, 7769.
- (20) Lobayan, R. M.; Boichichio, R. C.; Torre, A.; Lain, L. *J. Chem. Theory Comput.* **2009**, *5*, 2030 and references therein.
- (21) Lobayan, R. M.; Boichichio, R. C.; Lain, L.; Torre, A. *J. Chem. Phys.* **2005**, *123*, 144116.
- (22) Lobayan, R. M.; Boichichio, R. C.; Lain, L.; Torre, A. *J. Phys. Chem. A* **2007**, *111*, 3166.
- (23) Fox, M. A.; Wade, K. *Pure Appl. Chem.* **2003**, *75*, 1315 and references therein.
- (24) von Ragué Schleyer, P.; Subramanian, G.; Dransfeld, A. *J. Am. Chem. Soc.* **1996**, *118*, 9988.
- (25) Torre, A.; Lain, L.; Boichichio, R.; Ponec, R. *J. Comput. Chem.* **1999**, *20*, 1085.
- (26) Lobayan, R. M.; Alcoba, D. R.; Boichichio, R. C.; Torre, A.; Lain, L. *J. Phys. Chem. A* **2010**, *114*, 1200.
- (27) Luzanov, A. V.; Prezhdo, O. V. *Mol. Phys.* **2007**, *105*, 2879.
- (28) Lain, L.; Torre, A.; Alcoba, D. R.; Boichichio, R. C. *Chem. Phys. Lett.* **2009**, *476*, 101.
- (29) Torre, A.; Lain, L.; Boichichio, R. *J. Phys. Chem. A* **2003**, *107*, 127.
- (30) Lain, L.; Torre, A.; Boichichio, R. *J. Phys. Chem. A* **2004**, *108*, 4132.
- (31) Cioslowski, J.; Mixon, S. T. *J. Am. Chem. Soc.* **1991**, *113*, 4142.
- (32) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision C.02; Gaussian, Inc.: Wallingford, CT, USA, 2004.
- (33) Biegler-König, F. W.; Bader, R. F. W.; Tang, T. H. *J. Comput. Chem.* **1982**, *3*, 317.
- (34) Wade, K. *Nat. Chem.* **2009**, *1*, 92.
- (35) Popelier, P. L. A. *Coord. Chem. Rev.* **2000**, *197*, 169.
- (36) Boichichio, R.; Ponec, R.; Uhlik, P. *Inorg. Chem.* **1997**, *36*, 5363 and references therein.

Accurate Conformational Energy Differences of Carbohydrates: A Complete Basis Set Extrapolation

Gábor I. Csonka^{*,†} and Jakub Kaminsky[‡][†]Department of Inorganic and Analytical Chemistry, Budapest University of Technology, Szent Gellért tér 4, Budapest, H-1521 Hungary[‡]Department of Molecular Spectroscopy, Institute of Organic Chemistry and Biochemistry, Flemingovo nam. 2, 166 10 Prague, Czech Republic Supporting Information

ABSTRACT: Correlated ab initio wave function calculations have been performed, using nonrelativistic frozen core MP2 complete basis set extrapolation model chemistry. The calculations have been made for three test sets of gas-phase saccharide conformations to provide reference values for their relative energies. The remaining correlation effects are estimated from frozen core coupled-cluster singles and doubles [CCSD(T)] calculations. The test sets consist of 15 conformers of α - and β -D-allopyranose, 15 of 3,6-anhydro-4-O-methyl-D-galactitol, and four of β -D-glucopyranose. For each set, conformational energies varied by about 7 kcal/mol. These benchmark quality relative conformational energies are used to re-evaluate the performance of the best density functional methods for conformational analyses of saccharides. Our results show that the B3PW91 and PBE0 relative energies are systematically better than the B3LYP and M05-2X results. Overall, the functionals based on the exact constraints perform better for the relative energies of monosaccharide conformers than the empirically fitted functionals.

1. INTRODUCTION

This paper is a continuation of our recent study about the performance of various model chemistries on carbohydrate conformations.¹ We use three test sets that sample the lowest 6–7 kcal/mol energy range of conformation space. The AnGol15 test set contains 15 conformers of 3,6-anhydro-4-O-methyl-D-galactitol (cf. Figure 1). The GLC4 test set contains two low energy ⁴C₁ chair and two higher energy ¹C₄ chair forms of β -D-glucopyranose. The ALL15 test set contains 13 ⁴C₁ structures (8 α - and 5 β -anomers, with *gg*, *gt*, and *tg* hydroxymethyl rotamers) and 2 ¹C₄ conformers of α - and β -D-allopyranose.

In our previous paper¹ we tested nonempirical functionals, such as the local spin density approximation (LSDA), generalized gradient approximation GGA (e.g., Perdew–Burke–Ernzerhof, PBE)², and meta-GGA (e.g., Tao–Perdew–Staroverov–Scuseria, TPSS).³ We also tested semiempirical global hybrid functionals like PBE0,^{4,5} B3LYP,⁶ and B3PW91^{7,8} and a many-fit-parameter empirical hybrid functional, M05-2X.^{9,10} Our results showed a good performance of M05-2X methods for the ALL15 and AnGol15 test sets. However, for the GLC4 test set, the PBE and several other functionals performed better than the M05-2X functional.¹

Because the best density functionals might reach a considerable accuracy, the uncertainties in the reference relative energies for the conformational space of carbohydrates make the evaluation and ranking of the approximate methods uncertain. Monosaccharides are relatively large molecules (containing almost 100 electrons). Thus obtaining accurate relative energies requires a very large computational effort that uses a correlated wave function.

Calculations using the canonical MP2 method ($O(N)^5$ scaling of computer time with the size N) are such an effort. The slow convergence of the relative energies with the cardinal number of

the basis set also worsens the problem. An earlier canonical MP2 study¹¹ found that for alanine octapeptides, the basis set errors exceeded 4 kcal/mol when the augmented triple- ζ basis set was used.

The pseudospectral local MP2 (LMP2) approximation^{12,13} of the canonical MP2 is considerably faster than MP2 and approaches linear scaling for large systems. Another possible advantage is that the LMP2 results are less affected by the basis set size. Even augmented double- ζ basis sets yielded reasonably accurate relative energies for peptides if the domain selection is based on the natural population analysis and natural localized molecular orbitals.^{11,14} It was observed that this method yields more stable domains with respect to the basis set and geometry of peptides than the conventional method (for details see ref 11). However, LMP2 neglects dispersion-relevant double excitation terms, and a consistent domain selection is critically important for good relative energies of different conformers. We have found¹ that the LMP2/cc-pVTZ(-f) and MP2/aug-cc-pVTZ model chemistries give somewhat similar relative energies for the conformers in the GLC4 and ALL15 test sets. However, the agreement between LMP2 and MP2 results was considerably worse for larger basis sets and for the AnGol15 test set, possibly due to problems with the applied LMP2 domain selection method. In this paper we check whether this divergence between the LMP2 and MP2 methods persists at the basis set limit.

Finally, as the MP2 method misses a large part of the correlation energy, the more expensive CCSD(T) calculations ($O(N)^7$ scaling of computer time with the size N) might be necessary for improved relative energies. Recently Goerigk and

Received: January 4, 2011

Published: March 11, 2011

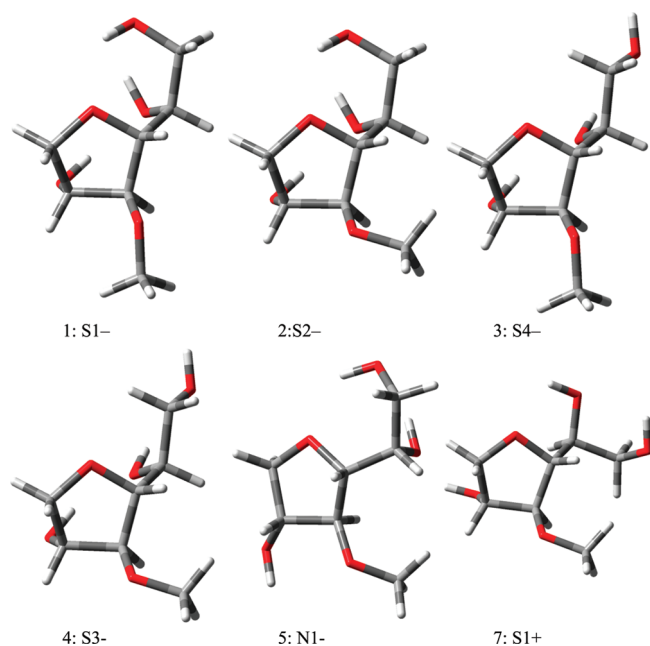


Figure 1. The five lowest energy and the seventh S1+ conformers of the AnGol15 conformational space.

Grimme¹⁵ performed high-level calculations for the AnGol15 and the GLC4 test sets (called the SCONF subset of the GMTKN24 test set). The MP2/complete basis set (CBS) limit was estimated from aug-cc-pVDZ and TZ energies. The differences between CCSD(T) and MP2 correlation energies for the AnGol15 test set were estimated with the cc-pVDZ basis set. The reference values for the GLC4 set were provided by one of us in a private communication (MP2/CBS with aug-cc-pVTZ and aug-cc-pVQZ; difference between CCSD(T)/cc-pVTZ and MP2/cc-pVTZ).¹⁵

In the current study we go further and perform calculations up to the MP2/cc-pV5Z model chemistry using the optimized B3LYP/6-31+G(d,p) geometries. For more accurate estimation of the relative energies we also include CCSD(T) results for the GLC4 test set. This allows the accuracy of our earlier reference data to be checked and we can replace the less accurate reference values.

We observed in our earlier study that our assessment of the performance of a given density functional theory (DFT) approximation might depend critically on the choice of the reference conformer (see Section 4.1). In this study we shall show how the choice of the reference conformer influences the accuracy of the CBS extrapolation. We shall select the conformer pairs for which the CBS extrapolation is particularly accurate, and we shall use these accurately known relative energies for the evaluation of the functionals. Given the new reference data, we re-evaluate the best DFT functionals.

2. METHODS

2.1. Complete Basis Set Extrapolation. We have performed frozen core MP2 calculations with cc-pVXZ ($X = 3-5$) and aug-cc-pVXZ ($X = 3-4$) correlation-consistent basis sets denoted as VXZ and AVXZ, respectively, in the remainder of the article, where X is the cardinal number of the correlation-consistent basis set. As these basis sets are expensive, we also tested a simplified AV3Z basis set. The aV3Z(-df) basis set contains no diffuse and d functions for

H atoms and no f functions for C and O atoms. All calculations were performed with MOLPRO.¹⁶

The HF and the MP2 correlation complete basis set extrapolations were done separately. For a given conformer the MP2 total energy is composed from the HF and the MP2 correlation energy correction:

$$E_{\text{MP2}}(i) = E_{\text{HF}}(i) + E_{\text{MP2corr}}(i) \quad (1)$$

For two-point extrapolations we use the following simple formula:

$$E_{\text{m}}[X-1, X] = E_{\text{m}}[X] + C_{\text{m}}[X-1, X](E_{\text{m}}[X] - E_{\text{m}}[X-1]) \quad (2)$$

where $C_{\text{m}}[X-1, X]$ is the extrapolation coefficient for correlation-consistent basis sets, and m is the method, HF or MP2 correction for correlation. Note that different forms exist for two-point extrapolation, but all these formulas give the same $C_{\text{m}}[X-1, X]$ coefficient after fitting to the same reference data.

2.2. Extrapolation of the HF Energy. We do not discuss the various (inverse power law, exponential) forms of the fit equations here; more details are in ref 17. For the AV[3,4]Z extrapolation, the following $C_{\text{HF}}[X-1, X]$ coefficient can be derived from the $A + B/X^\alpha$ form:

$$C_{\text{HF}}[3, 4] = \frac{1}{\left(\frac{4}{3}\right)^\alpha - 1} = 0.274, \quad \text{rmsd} = 0.13 \text{ kcal/mol} \quad (3)$$

where $\alpha = 5.34$ is an effective decay exponent obtained by minimizing the root-mean-square deviation (rmsd) from the numerical HF $E_{\text{m}}[\infty]$ for the test set proposed by Jensen.¹⁸

Karton and Martin¹⁷ have found that the $A + B/X^5$ extrapolation used in W1 theory gives reasonable results for AV[3,4]Z extrapolation. (Note that the fitted $\alpha = 5.34$ parameter in eq 3 is close to 5). But the $A + B/X^5$ extrapolation gives poor results for AV[4, 5]Z and AV[5, 6]Z extrapolation. Those results are worse than the simple AV5Z and AV6Z results, respectively.

For the more accurate AV[4,5]Z extrapolation, a different optimized coefficient was found:¹⁷

$$C_{\text{HF}}[4, 5] = \frac{1}{\left(\frac{5}{4}\right)^{8.74} - 1} = 0.166, \quad \text{rmsd} = 0.08 \text{ kcal/mol} \quad (4)$$

It was also observed¹⁷ that the conventional three-point geometric extrapolation with the AV[3,4,5]Z basis sets gives inferior results compared to the simple AV5Z results.

The aug-cc-pV($n + d$)Z basis sets¹⁹ contain an extra high-exponent d function for second-row atoms, in order to recover "inner polarization" effects.²⁰ For the aug-cc-pV($n + d$)Z basis sets ($n \geq 4$), a very similar $C_{\text{HF}}[4, 5]$ was found:¹⁷

$$C_{\text{HF}}[4, 5 + d] = \frac{1}{\frac{5e^{9(\sqrt{5}-2)}}{6} - 1} = 0.167 \quad (5)$$

The optimized coefficients in eqs 3–5 might show some transferability, as comparison of eqs 4 and 5 shows some basis set independence. But in general such coefficients should be reoptimized for different test and basis sets, leading to various empirical coefficients.

2.3. Extrapolation of the Correlation Energy. For frozen core MP2 or coupled cluster (CC) correlation energy, it was assumed that the basis set convergence follows an X^{-3} law with respect to the cardinal number X of the correlation-consistent basis sets.²¹ The two-point CBS MP2 correction for correlation can be derived from the form $A + B/X$:³

$$C_{\text{MP2corr}}[3, 4] = \frac{1}{\left(\frac{4}{3}\right)^3 - 1} \approx 0.730 \quad \text{and}$$

$$C_{\text{MP2corr}}[4, 5] = \frac{1}{\left(\frac{5}{4}\right)^3 - 1} \approx 1.049 \quad (6)$$

Notice that L^{-3} and L^{-5} laws were found for the unlike- and the like-spin pairs for MP2 correlation energy, where L is the largest angular momentum in the basis set.²² Even in these cases the basis set convergence follows an X^{-3} law. However, for the He–He interatomic potential, separate extrapolation of singlet and triplet pairs improves the results.²³ Notice also that the extrapolation formula in eq 2 is a linear combination of two energies. Consequently optimized geometries and transition structures can be obtained readily from linear combinations of gradients and Hessians.^{24,25} The higher level correlation is estimated from the difference of MP2 and CCSD(T) relative energies calculated with a moderately large AV3Z(–df) basis set. This simplification is necessary as the CCSD(T)/AV3Z calculations are prohibitively expensive for monosaccharides. It is assumed that this procedure gives reasonable CCSD(T)/CBS relative energy estimations.

3. RELATIVE ENERGIES

The relative energies are defined as the difference between the energy of the i^{th} conformer and the energy of an r reference conformer ($i \neq r$) using the given model chemistry:

$$\Delta E_{\text{model}}(i, r) = E_{\text{model}}(i) - E_{\text{model}}(r) \quad (7)$$

Note that this way the large total energy terms translate into much smaller reference relative energy terms.

The relative energy difference for the i^{th} conformer between two model chemistries, modelA and modelB is:

$$\Delta \Delta E_{\text{modelA-modelB}}(i, r) = \Delta E_{\text{modelA}}(i, r) - \Delta E_{\text{modelB}}(i, r) \quad (8)$$

where the difference between the two models can be in the basis set or in the treatment of electron correlation. For the present paper the relative energy difference between correlated (MP2 or CCSD(T)) and HF methods is particularly important and denoted as $\Delta E_{\text{MP2corr}}(i, r)$. To obtain MP2/CBS relative energies, we sum CBS estimations of the HF and MP2corr:

$$\Delta E_{\text{MP2/CBS}}(i, r) = \Delta E_{\text{MP2corr/CBS}}(i, r) + \Delta E_{\text{HF/CBS}}(i, r) \quad (9)$$

The model chemistry and reference conformer dependent mean deviation (MD) is defined as

$$\text{MD}_{\text{model A-model B}}(r) = \frac{1}{n-1} \sum_{i=1}^n \Delta \Delta E_{\text{model A-model B}}(i, r) \quad (10)$$

and the model and reference conformer dependent mean absolute deviation (MAD) is defined as:

$$\text{MAD}_{\text{model A-model B}}(r) = \frac{1}{n-1} \sum_{i=1}^n |\Delta \Delta E_{\text{model A-model B}}(i, r)| \quad (11)$$

Within a given test set of conformers, the mean deviation and the mean deviation between the two compared models depend on the choice of the reference conformer. This makes the evaluation and comparison of model chemistries noninvariant under the choice of the reference conformer. The range of the relative difference (RRD) = $\max \Delta \Delta E - \min \Delta \Delta E$ can be used as an invariant measure of the model chemistry dependent but reference conformer independent relative energies, where $\max \Delta \Delta E$ is the most positive and $\min \Delta \Delta E$ is the most negative relative energy difference.

For specific groups of conformers, accurate relative energies can be obtained at a considerably reduced cost, if one or more of the following conditions are fulfilled:

- (1) The basis set dependence of the relative HF energy is less than 0.1 kcal/mol: $|\Delta E_{\text{HF/V4Z}}(i, r) - \Delta E_{\text{HF/V3Z}}(i, r)| < 0.1$ kcal/mol.
- (2) The relative correlation energy is small: $|\Delta E_{\text{MP2corr/V3Z}}(i, r)| < 0.05$ kcal/mol.
- (3) The basis set dependence of the relative correlation energy is less than 0.1 kcal/mol: $|\Delta E_{\text{MP2corr/V4Z}}(i, r) - \Delta E_{\text{MP2corr/V3Z}}(i, r)| < 0.1$ kcal/mol.

The consequence of fulfilling conditions 1 and 2 is that the MP2/CBS relative energies can be estimated from HF/V3Z results with a small error. This is the basis of the earlier observations that HF relative energies are surprisingly good for several test sets. But this good performance is occasional and based on special structural similarity, as we shall show in this paper. Fulfillment of condition 3 alone makes the MP2/V3Z results a good estimation of the MP2/CBS relative energies.

4. RESULTS AND DISCUSSION

4.1. 3,6-Anhydro-4-O-methyl-D-galactitol Conformers. The schematic representations of the torsion angles in the 3,6-anhydro-4-O-methyl-D-galactitol (AnGol) and the full conformational space can be found in refs 1 and 26. AnGol has a flexible five-membered ring and six exocyclic torsion angles. The two main stable conformations of the five-membered ring were denoted as North (N) (Cremer–Pople puckering²⁷ parameter $\phi \approx 250-320^\circ$) and South (S) ($\phi \approx 120^\circ$). The 15 conformers numbered from 1 to 15 were designated as S1–, S2–, S4–, S3–, N1–, N2–, S1+, S2+, N3–, N1+, S4+, S3+, N3+, N2+, and N4+, respectively in refs 1 and 26. The positions of the exocyclic torsions of six conformers are shown in Figure 1. These torsions are explained in detail in the Supporting Information. This conformational space differs from the typical monosaccharide conformational space of allopuranose or glucopyranose, as also shown in the Supporting Information.

Traditionally the most stable conformer is selected as reference. Figure 2 shows the dependence of the HF relative energies on the cardinal number X of the VXZ basis sets if S1– is the reference conformer (the relevant total energies are shown in the Supporting Information). Condition 1 is fulfilled for the relative HF energies of the first two conformers, S1– and S2–, and the

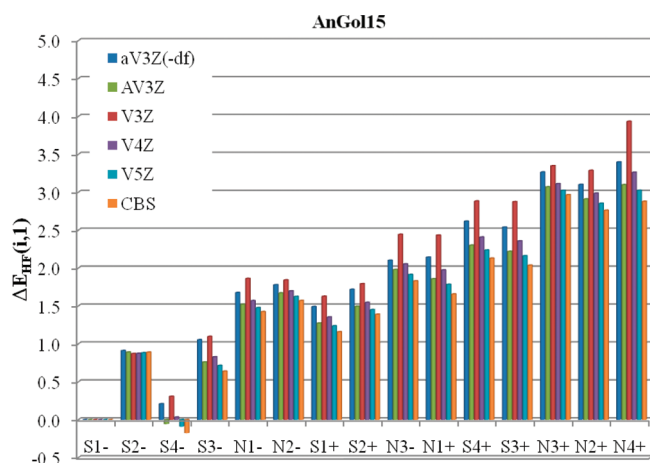


Figure 2. The HF relative energies, $\Delta E_{\text{HF}}(i, 1)$ (kcal/mol) of the members of the AnGol15 test set, calculated with basis sets with cardinal numbers from 3 to 5, compared to the most stable S1– conformer (no. 1).

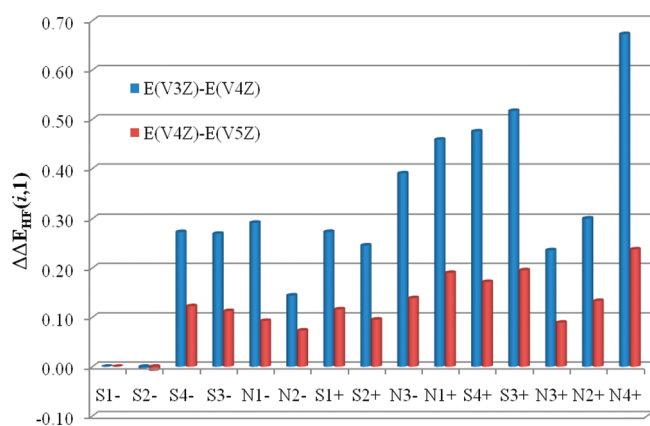


Figure 3. The HF relative energy differences, $\Delta \Delta E_{\text{HF}}(i, 1)$ (kcal/mol), between the elements of the AnGol15 test set, compared to the most stable S1– conformer (no. 1). $E(\text{VXZ}) - E(\text{VX} + 1\text{Z})$ means the difference between the relative energies for cardinal number of X.

HF/V3Z result is practically converged. The figure also shows that this is not true for the other 13 conformers.

Figure 3 shows how the HF relative energy differences, $\Delta \Delta E_{\text{HF}}(i, 1)$ (kcal/mol) change with increase of the cardinal number of the basis set by one, if the S1– conformer is the reference. This figure shows how small is the basis set dependence of the HF relative energies for S1– and S2– conformers. However, this is not true for the other conformers; the largest difference can be observed for N4+ conformer. The figure also shows how much better the V4Z basis set is than the V3Z (RRD = 0.22 vs 0.68 kcal/mol, respectively).

Figure 4 shows how the basis set convergence changes if we choose the third (S4–) conformer as reference. The relative HF energies of S4–, S3–, and S1+ conformers converge parallel with the increase of the basis set, and the relative HF energies N1–, S2+, N3+, and N2+ conformers behave very similarly within 0.1 kcal/mol. The sets {S1–, S2–} and {S4–, S3–, S1+, N1–, S2+, N3+, N2+} fulfill criterion 1. Thus even the relative energies calculated with V3Z basis set are practically converged. A similar observation can be made for the {N1+, S4+, S3+} set.

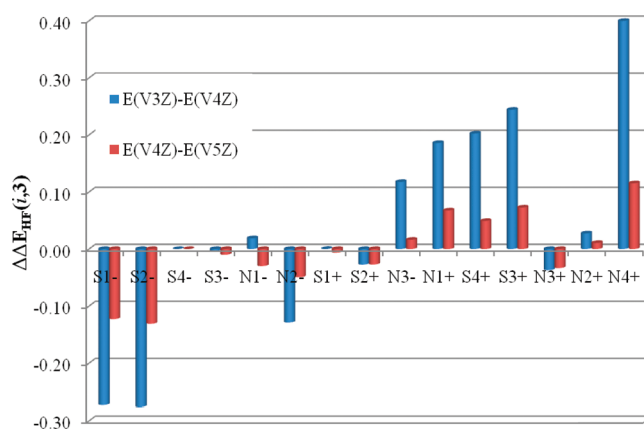


Figure 4. The HF relative energy differences, $\Delta \Delta E_{\text{HF}}(i, 3)$ (kcal/mol), between the elements of the AnGol15 test set, compared to the S4– conformer (no. 3).

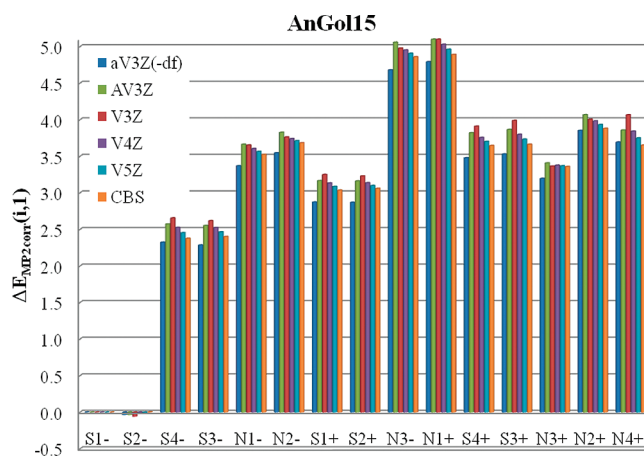


Figure 5. The MP2 relative correlation energies, $\Delta E_{\text{MP2corr}}(i, 1)$ (kcal/mol), of the elements of the AnGol15 test set, calculated with basis sets with cardinal numbers from 3 to 5, compared to the most stable S1– conformer (no. 1).

Figure 5 shows the AVXZ and VXZ ($X = 3-5$) basis set dependence of the relative MP2 correlation energies compared to the lowest energy S1– conformer. Comparison of the Figures 2 and 5 shows the important contribution of the MP2 correlation energy to the relative energies. The HF/V5Z relative energy differences span a range of 3 kcal/mol, while the MP2corr/V5Z energy differences span almost 5 kcal/mol. Another, somewhat surprising observation is that the HF relative energies converge in bigger steps than that of the MP2corr relative energies. Consequently for several conformers the larger part of the error of MP2/V3Z relative energy comes from the HF error. So the convergence of the HF relative energies is particularly important for these conformers. This again depends on the choice of the reference conformer (cf. Figures 3, 4, 6, and 7). The results show that there is almost no difference between the MP2 relative correlation energies of S1– and S2– conformers (conditions 1 and 2 are fulfilled). This makes the $\Delta E_{\text{MP2/CBS}}(2, 1) \approx \Delta E_{\text{HF/V3Z}}(2, 1)$. Thus for these two conformers the inexpensive HF/V3Z relative energy (0.87 kcal/mol) is close to the very expensive MP2/CBS[4, 5] relative energy (0.88 kcal/mol). Notice that the CCSD(T)/CBS relative energy is 0.83 kcal/mol,¹⁵ thus $\Delta E_{\text{CCSD(T)/CBS}}(2, 1) \approx \Delta E_{\text{HF/V3Z}}(2, 1)$, with the negligible error of 0.05 kcal/mol.

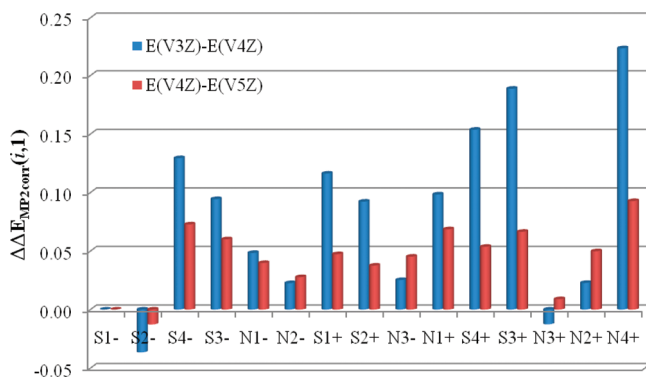


Figure 6. The MP2 correlation relative energy differences, $\Delta\Delta E_{\text{MP2corr}}(i, 1)$ (kcal/mol), between the elements of the AnGol15 test set, compared to the most stable S1− conformer (no. 1). $E(\text{VXZ}) - E(\text{VX} + 1\text{Z})$ means the difference between the relative energies for cardinal number of X.

Figure 6 shows that the relative MP2 correlation energy is practically basis set independent for S1− and N3+ (1 and 13) conformers. Thus condition 3 is fulfilled, and this makes the $\Delta E_{\text{MP2corr/CBS}}(13, 1) = \Delta E_{\text{MP2corr/V3Z}}(13, 1) = 3.36$ kcal/mol. As condition 1 is not valid, we have to calculate the HF/CBS energy: $\Delta E_{\text{MP2/CBS}}(13, 1) \approx \Delta E_{\text{MP2corr/V3Z}}(13, 1) + \Delta E_{\text{HF/CBS}}(13, 1)$. Figure 6 also shows that condition 3 is also valid for N1−, N2−, N3−, and N2+ (5, 6, 9, and 14) conformers and $|\Delta E_{\text{MP2corr/CBS}}(i, 1) - \Delta E_{\text{MP2corr/V3Z}}(i, 1)| < 0.13$ kcal/mol for these conformers. This condition makes the MP2corr/CBS extrapolation quite reliable.

Figure 7 shows that the relative MP2 correlation energies of the S4−, S3−, S1+, S2+, N1+, and S4+ (3, 4, 7, 8, 10, and 11) conformations also fulfill condition 3, making MP2corr/CBS extrapolation very accurate within this subset.

Table 1 summarizes the MP2 and CCSD(T) results and statistics for the relative energies compared to S1− conformer. We use the CCSD(T)/CBS results¹⁵ as reference. The MP2/CBS[3, 4, 5] uses the conventional three point geometric extrapolation formula for HF basis set limit. The MP2/CBS[4, 5] uses the HF limits obtained from eq 4. Both extrapolations use the two point MP2corr/CBS[4, 5] (cf. eq 6). The MP2/CBS[4, 5] results are closer to the MP2/V5Z and CCSD(T)/CBS results. We have also performed an MP2/CBS[A3, A4] extrapolation using eqs 3 and 6. This extrapolation uses AV3Z and AV4Z basis sets, and it gives particularly good agreement with the MP2/V5Z results. The economical 6-311+G(d,p) basis set used in earlier studies²⁶ gives the largest basis set error. The simplified aV3Z(−df) basis set shows a particularly good performance due to the compensation of the HF and MP2 errors (cf. Figures 2 and 5); however, such an error compensation is unreliable as we shall show in Section 4.3. The diffuse functions added to the triple- ζ basis sets of the oxygen atoms are essential for the correct description of the electron densities around the oxygen atoms (lone pairs) and the O–H···O interactions. However, on the hydrogen atoms the diffuse functions have a negligible effect on the relative energies.

The results show that for the AnGol15 test set, the LMP2 and MP2 results diverge at the basis set limit (cf. Table S3 in the Supporting Information). This is in agreement with our previous observation of a considerable difference between the MP2 and LMP2/AV3Z relative energies for these conformers.¹ Similar problems were noticed with the LMP2 implementation in the JAGUAR program.²⁸

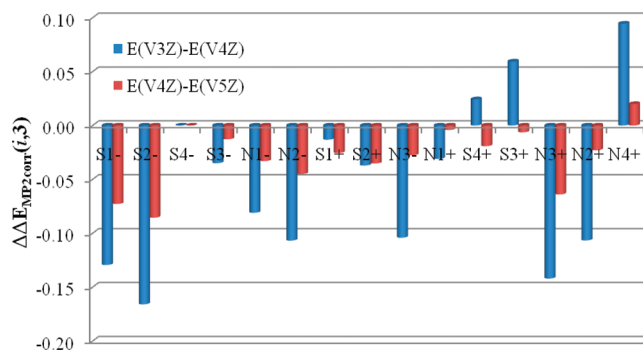


Figure 7. The MP2 correlation relative energy differences, $\Delta\Delta E_{\text{MP2corr}}(i, 3)$ (kcal/mol), between the elements of the AnGol15 test set, compared to the S4− conformer (no. 3). $E(\text{VXZ}) - E(\text{VX} + 1\text{Z})$ means the difference between the relative energies for cardinal number of X.

Finally we summarize the most accurately known relative energies in Table 2 for various reference conformers. Analysis of the results shows the excellent performance of the smallest aV3Z(−df) basis set for these conformer pairs and the almost perfect agreement between MP2/V5Z, CBS[A3, A4], and CBS[4, 5] results (MAD = 0.03 kcal/mol). The MP2 and CCSD(T) results also agree well statistically (MAD = 0.1 kcal/mol). The largest CCSD(T) correction (about 0.2 kcal/mol) to the MP2 can be found for the relative energies of {13, 1}, {10, 3} and {13, 14} conformer pairs.

4.2. β -D-glucopyranose Conformers. The GLC4 test set is composed of two low-energy ${}^4\text{C}_1$ conformers (1 and 2) and two ${}^1\text{C}_4$ conformers (3 and 4) of β -D-glucopyranose, as shown in Figure 8. These four conformers were used for testing KS-DFT functionals before.^{29,30} In our previous work,¹ we used the MP2/aV3Z(−df)//B3LYP/6-31+G(d,p) model chemistry as a reference (cf. Table 3).

The GLC4 test set is an interesting example for a situation in which the less stable ${}^1\text{C}_4$ conformer has stronger stabilizing electron correlation effects than the most stable ${}^4\text{C}_1$ conformer. This situation is very different from that of observed for the AnGol15 test set in which the electron correlation effects mostly destabilize the higher energy conformers (cf. Figure 5). If one unites the AnGol15 and GLC4 test sets, the opposite errors from incorrect treatment of the electron correlation might cancel, and the results might improve statistically but not in reality. We treat the two test sets separately.

It was observed earlier that the double- ζ polarized (DZP) basis set gives surprisingly good HF relative energies for ${}^1\text{C}_4$ and ${}^4\text{C}_1$ conformers of β -D-glucopyranose^{29,30} and very poor MP2 relative energies. From our results it is clear that the origin of these good HF/DZP results is a large, 6–7 kcal/mol basis set error that erroneously stabilizes the ${}^1\text{C}_4$ conformers. The opposite exchange–correlation effects might systematically compensate each other^{29,30} and help the semilocal DFT approximations to give reasonable results.

In Table 3 we show our highest level CCSD(T)/CBS estimations of the relative energies. These were calculated from MP2/CBS[4, 5] and CCSD(T) correction to the MP2/V3Z(−df) relative energies. This correction stabilizes the ${}^1\text{C}_4$ conformers by about 0.2–0.3 kcal/mol compared to the reference conformer. The results in Table 3 show again the surprisingly good performance of the least expensive MP2/aV3Z(−df) model chemistry (MAD = 0.2 kcal/

Table 1. Relative Energies of AnGol Conformers (kcal/mol), Compared to S1– Conformer Calculated with MP2, and CCSD(T) Methods, Various Basis Sets, and CBS Extrapolations^a

no.	conformer	MP2					CCSD(T)	
		6-311+G(d,p) ^b	aV3Z(-df) ^b	AV3Z ^c	VSZ ^c	CBS[3, 4, 5] ^d	CBS[4, 5] ^e	CBS ^f
1	S1–	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	S2–	1.05	0.87	0.85	0.87	0.87	0.88	0.83
3	S4–	2.11	2.53	2.52	2.36	2.19	2.27	2.60
4	S3–	3.07	3.34	3.31	3.18	3.03	3.10	3.37
5	N1–	5.15	5.05	5.18	5.04	4.96	4.98	4.87
6	N2–	5.51	5.32	5.50	5.34	5.23	5.29	5.18
7	S1+	4.23	4.36	4.44	4.32	4.19	4.25	4.47
8	S2+	4.54	4.59	4.66	4.55	4.45	4.49	4.68
9	N3–	6.91	6.78	7.04	6.82	6.70	6.75	6.69
10	N1+	6.90	6.93	7.00	6.75	6.54	6.64	6.75
11	S4+	6.05	6.09	6.12	5.94	5.78	5.85	6.08
12	S3+	5.99	6.07	6.09	5.89	5.71	5.79	6.05
13	N3+	6.78	6.46	6.48	6.39	6.33	6.37	6.17
14	N2+	7.05	6.95	6.98	6.79	6.63	6.71	6.75
15	N4+	7.10	7.09	6.96	6.77	6.55	6.64	6.71
MD		0.09	0.09	0.14	–0.01	–0.15	–0.08	
MAD		0.25	0.13	0.16	0.13	0.19	0.16	

^a Mean deviation, MD, and mean absolute deviation, MAD, compared to CCSD(T)/CBS; see eqs 9 and 10. ^b Ref 1. ^c This work. ^d This work: HF/CBS[3, 4, 5] + MP2/CBS[4, 5]. ^e This work: HF/CBS[4, 5] + MP2/CBS[4, 5]. ^f Ref 15.

mol). This confirms that our earlier reference¹ is relatively good, but it cannot distinguish between KS-DFT functionals with small errors. Our results show that the more expensive MP2/V3Z and AV3Z models perform worse (MAD = 1.28 and 0.27 kcal/mol, respectively). These results support our earlier observations about the particular importance of the diffuse functions on the oxygen atoms (cf. O–H···O interactions) and their unimportance for the hydrogen atoms. The CBS[A3, A4] limit results agree particularly well again with the almost converged VSZ results. The conventional three-point geometric extrapolation formula for HF combined with a two-point MP2corr/CBS[4, 5] extrapolation, annotated as CBS-[3, 4, 5] in Table 3, shows again worse agreement with the MP2/V3Z and CCSD(T)/CBS results than the CBS[4, 5] extrapolation. There is a slight disagreement between our new results and the previous CCSD(T)/CBS¹⁵ estimation of the relative energies (MAD = 0.07 kcal/mol) because here we used the MP2/CBS[4, 5] estimation that was not available earlier.

4.3. D-Allopyranose Conformers. For the ALL15 test set¹ we selected 15 optimized B3LYP/6-31+G(d,p) conformers out of 102 recently published α - and β -D-allopyranose conformers.³¹ The test set includes 13 ⁴C₁ conformers (8 α - and 5 β -anomers, with *gg*, *gt*, and *tg* hydroxymethyl rotamers), one α - and one β -¹C₄ conformer. The first, low-energy reference conformer is the ⁴C₁ α -D-allopyranose denoted as conformer 1 in Figure 9. Ref 1 gives a detailed description of the ring conformations, the anomers, and the exocyclic torsional angles. This test set is particularly suitable for demonstrating the importance of electron correlation for anomeric effects and ring conformations.

In our previous paper¹ we chose the MP2/aV3Z(-df) relative energies as reference for the ALL15 test set. The results in Table 4 show a reasonable performance of the MP2/aV3Z(-df) model chemistry (MAD = 0.2 kcal/mol, RRD = 0.8 kcal/mol). The error compensation of the aV3Z(-df) basis set is less

efficient for the ALL15 conformers than it was for the AnGol15 conformers. Because of this, our earlier reference¹ is not suitable to correctly judge the performance of the functionals with small 0.2–0.5 kcal/mol deviations. The MP2/AV3Z relative energies are better (MAD = 0.1 and RRD = 0.4 kcal/mol), and the MP2/V3Z relative energies are very well converged. Consequently the MP2/CBS[4, 5] relative energies are reliable within 0.1 kcal/mol.

Conditions 1 and 2 are valid for the subset of {1, 8, 15}; thus even the HF/V4Z model chemistry gives converged relative energies. For the subsets of the conformers {1, 4, 5, 7, 8, 10, 14, 15} and {2, 6, 9, 11, 12} the condition 3 is valid. For these subsets even the MP2/AV3Z relative energies are well converged (cf. Table 4).

According to the HF/CBS results the most stable conformer is 3 (β -⁴C₁) (cf. Figure 9) and $\Delta E_{\text{HF/CBS}}(3, 1) = -2.63$ kcal/mol. The HF method erroneously stabilizes the β anomers by about 2.8–3.3 kcal/mol (cf. conformers 2, 3, and 6 in Table 4; the HF energies can be found in the Supporting Information). The MP2 electron correlation effects stabilize the α anomers and make conformer 1 (α -⁴C₁) the most stable (cf. Table 4). This is another example of the stabilizing correlation effects (cf. the GLC4 test set).

5. EVALUATION OF THE DENSITY FUNCTIONAL RESULTS

The most stable conformers of the AnGol15, GLC4, and ALL15 conformational space are special in the gas phase. In these conformers the O–H···O interactions show stabilizing cooperative effects. Overestimation of these stabilizing effects makes the first conformer too stable compared to the other conformers and leads to too-large relative energies. Underestimation of these stabilization effects leads to too-small relative energies and thus negative deviations from the accurate relative energies. The GLC4 test set is an interesting counter example: the less

Table 2. Well Converged Relative Energies (kcal/mol) of Selected AnGol Conformers (Conformer *i*), Compared to Various Reference Conformers Calculated with MP2, and CCSD(T) Methods and Various Basis Sets and CBS Extrapolations^a

conformers	MP2				CCSD(T)	
	<i>i</i>	reference	aV3Z(-df) ^b	VSZ ^b	CBS[A3, A4] ^c	CBS[4, 5] ^d
2	1	0.87	0.87	0.88	0.88	0.83
5	1	5.05	5.04	5.03	4.98	4.87
6	1	5.32	5.34	5.35	5.29	5.18
9	1	6.78	6.82	6.82	6.75	6.69
13	1	6.46	6.39	6.38	6.37	6.17
14	1	6.95	6.79	6.81	6.71	6.75
5	2	4.17	4.17	4.16	4.11	4.04
6	2	4.45	4.46	4.47	4.42	4.34
9	2	5.91	5.95	5.95	5.87	5.86
13	2	5.59	5.52	5.51	5.49	5.34
14	2	6.08	5.91	5.94	5.83	5.92
4	3	0.81	0.81	0.83	0.83	0.77
7	3	1.83	1.96	2.02	1.99	1.87
8	3	2.05	2.19	2.26	2.23	2.08
10	3	4.40	4.38	4.40	4.38	4.15
11	3	3.56	3.57	3.62	3.59	3.48
7	4	1.02	1.14	1.20	1.16	1.11
8	4	1.25	1.37	1.44	1.40	1.31
10	4	3.59	3.57	3.57	3.55	3.38
11	4	2.76	2.76	2.80	2.76	2.72
6	5	0.28	0.30	0.32	0.31	0.31
9	5	1.74	1.78	1.79	1.77	1.82
13	5	1.42	1.35	1.35	1.38	1.30
14	5	1.91	1.74	1.78	1.73	1.88
9	6	1.46	1.49	1.47	1.46	1.51
13	6	1.14	1.05	1.03	1.07	0.99
14	6	1.63	1.45	1.46	1.42	1.58
8	7	0.22	0.23	0.24	0.24	0.20
10	7	2.57	2.42	2.38	2.39	2.27
11	7	1.73	1.62	1.60	1.60	1.61
10	8	2.35	2.20	2.14	2.15	2.07
11	8	1.51	1.39	1.36	1.36	1.41
13	9	-0.32	-0.43	-0.44	-0.38	-0.52
14	9	0.17	-0.04	-0.01	-0.04	0.06
11	10	-0.84	-0.81	-0.78	-0.79	-0.67
14	13	0.49	0.40	0.43	0.34	0.58
MD		0.09	0.05	0.06	0.04	
MAD		0.12	0.10	0.10	0.10	

^aMean deviation, MD, and mean absolute deviation, MAD, compared to CCSD(T)/CBS; see eqs 9 and 10. ^bThis work. ^cThis work: HF/CBS[A3, A4] + MP2/CBS[A3, A4], where A means augmented basis set. ^dThis work: HF/CBS[4, 5] + MP2/CBS[4, 5]. ^eRef 15.

stable ¹C₄ conformer has stronger stabilizing electron correlation effects than the most stable ⁴C₁ conformer.

Our previous study showed¹ that LSDA strongly overestimates and the HF method strongly underestimates these stabilization effects. The GGAs and meta-GGAs improve on the

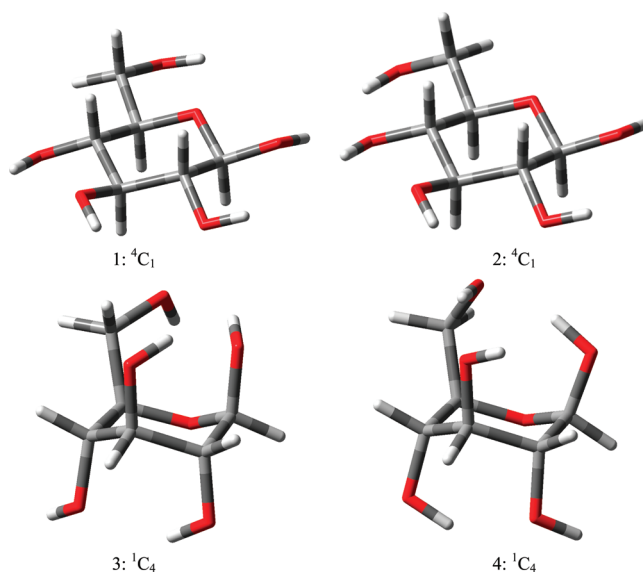


Figure 8. The four conformers of the GLC4 test set.

LSDA, but the relative conformational energies might remain statistically too large (mean deviation is smaller but positive) for the AnGol15 and ALL15 test sets.¹ For such (meta-)GGA functionals an empirical dispersion correction does not improve the results. Global mixing of the (meta-) GGAs with exact exchange might further improve the results by reducing overestimation of the relative energies, thus shifting the mean deviation close to zero. A sufficiently large exact exchange fraction might even lead to underestimation of the stabilizing correlation effects. Such errors of the global hybrid functionals can be efficiently corrected by a simple and quick empirical dispersion correction, as will be discussed later in this section.

To illustrate these tendencies, Figure 10 shows the AnGol15 $\Delta\Delta E_{\text{modelA-modelB}}(i, 1)$ relative energy differences for modelA = M05-2X, B3LYP, B3PW91, estimated CCSD(T)/CBS, PBE, and TPSS and for modelB = MP2/CBS[4, 5]. All the DFT results are from ref 1, and the estimated CCSD(T)/CBS results are from ref 15. In the DFT calculations the 6-311+G(d,p) basis set and the B3LYP/6-31+G(d,p) geometries were used. PBE and TPSS give excellent relative energies for the N1-, N2-, and N3- conformers but overestimate the relative energies of the S4-, S3-, S1+, and S2+ conformers. Interestingly the CCSD(T)/CBS, PBE, TPSS show qualitatively similar deviations from the MP2/CBS relative energies (cf. Figure 10). The opposite error occurs for B3LYP relative energies leading to serious underestimation of the relative energies of the N1-, N2-, and N3- conformers (cf. Figure 10). The mean absolute deviations and the ranges of the relative differences are similar for B3LYP, PBE, and TPSS (MAD = 0.5 kcal/mol, RRD = 1.5 kcal/mol). Notice that PBE and TPSS perform better than B3LYP if compared to CCSD(T)/CBS. The B3PW91 results are the best (MAD = 0.4 kcal/mol), but they show the same error pattern as the other three functionals (cf. the large RRD = 1.6 kcal/mol). The only functional that shows a different error pattern is the M05-2X (MAD = 0.4 kcal/mol, RRD = 1.0 kcal/mol). We have observed that different implementation of the M05-2X functional leads to worse results, and thus further study is required. The order of the functionals if compared to the CCSD(T)/CBS results is from best to worse: B3PW91, PBE = M05-2X, and

Table 3. Relative Energies of GLC4 β -D-Glucose Conformers (kcal/mol), Compared to The Most Stable 4C_1 Conformer Calculated with MP2, and CCSD(T) Methods, Various Basis Sets, and CBS Extrapolations^a

no.	ring conformation	MP2								CCSD(T)	
		aV3Z(-df) ^b	V3Z ^c	AV3Z ^c	V5Z ^c	CBS[3, 4] ^c	CBS[A3, A4] ^c	CBS[3, 4, 5] ^d	CBS[4, 5] ^e	CBS ^f	CBS ^c
1	4C_1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	4C_1	0.38	0.06	0.28	0.28	0.30	0.26	0.31	0.30	0.27	0.30
3	1C_4	5.76	4.30	5.68	6.14	6.25	6.12	6.40	6.33	5.92	6.01
4	1C_4	5.10	3.47	4.90	5.34	5.43	5.35	5.65	5.57	5.29	5.38
MD		-0.14	-1.28	-0.27	0.03	0.10	0.02	0.23	0.17	-0.07	
MAD		0.20	1.28	0.27	0.06	0.10	0.06	0.23	0.17	0.07	

^a Mean deviation, MD, and mean absolute deviation, MAD, compared to CCSD(T)/CBS; see eqs 9 and 10. ^b Ref 1. ^c This work. ^d This work: HF/CBS[3, 4, 5] + MP2/CBS[4, 5]. ^e This work: HF/CBS[4, 5] + MP2/CBS[4, 5]. ^f Ref 15.

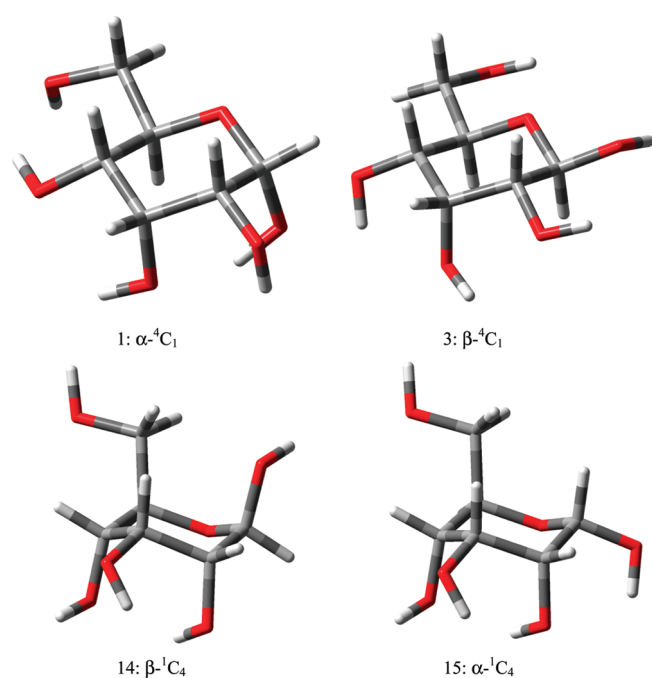


Figure 9. Four selected conformers of the ALL15 test set (conformers 1, 3, 14, and 15) representing the α and β anomers of 4C_1 and 1C_4 hexapyranose rings.

TPSS = B3LYP (MAD = 0.3, 0.4, 0.4, 0.6, and 0.6 kcal/mol, respectively).

We have also compared the DFT results to the particularly accurate relative energies shown in the Table 2. For these conformer pairs the PBE results show the best agreement with the reference (MAD = 0.3 kcal/mol), closely followed by TPSS and M05-2X results. The B3LYP results agree the worst (MAD = 0.6 kcal/mol).

For the three relative energies of the GLC4 test set, the global hybrid PBE0 and B3PW91 functionals perform equally well (MAD = 0.3 kcal/mol), while B3LYP is somewhat worse (MAD = 0.5 kcal/mol). Notice that the large overbinding basis error in the HF/DZP results helps to obtain reasonable results. The M05-2X performs the worst (MAD = 0.8 kcal/mol). In all calculations the 6-311+G(d,p) basis set and the optimized B3LYP/6-31G(d) geometries were used.¹

Table 4. Relative Energies of ALL15 conformers (kcal/mol), Compared to The Most Stable α - 4C_1 Conformer Calculated with MP2 Method, and Various Basis Sets and CBS Extrapolations^a

no.	conformer	MP2					
		aV3Z(-df)	AV3Z	V5Z	CBS[A3, A4]	CBS[3, 4, 5]	CBS[4, 5]
1	α - 4C_1	0.00	0.00	0.00	0.00	0.00	0.00
2	β - 4C_1	0.56	0.65	0.51	0.43	0.43	0.46
3	β - 4C_1	0.84	0.88	0.72	0.63	0.60	0.64
4	α - 4C_1	0.48	0.59	0.62	0.52	0.58	0.59
5	α - 4C_1	0.30	0.47	0.51	0.42	0.50	0.50
6	β - 4C_1	1.11	1.10	0.96	0.94	0.88	0.91
7	α - 4C_1	0.44	0.55	0.55	0.46	0.53	0.53
8	α - 4C_1	1.55	1.62	1.67	1.56	1.70	1.68
9	α - 4C_1	2.01	1.95	1.89	1.80	1.88	1.89
10	α - 4C_1	2.28	2.47	2.54	2.49	2.53	2.54
11	β - 4C_1	3.12	2.93	2.78	2.74	2.66	2.70
12	α - 4C_1	2.62	2.78	2.82	2.71	2.71	2.75
13	β - 4C_1	4.30	4.36	4.24	4.14	4.08	4.14
14	β - 1C_4	4.28	4.51	4.65	4.59	4.64	4.64
15	α - 1C_4	4.93	5.02	5.09	5.01	5.04	5.06
MD		-0.01	0.06	0.04	-0.04	-0.02	
MAD		0.18	0.11	0.04	0.05	0.02	

^a Mean deviation, MD, and mean absolute deviation, MAD, compared to CCSD(T)/CBS; see eqs 9 and 10.

Analysis of the earlier results for the SCONF test set¹⁵ (AnGol15 + GLC4) shows that in agreement with our results, the PBE and TPSS perform considerably better than the B3LYP functional. Our new results show that B3PW91 and the PBE0 functionals perform even better than PBE. A posteriori empirical dispersion correction (DFT-D)³² of the already too large PBE and TPSS relative energies deteriorates the results.¹⁵ The B3LYP relative energies that underestimate the intramolecular stabilization effects might be improved by the DFT-D correction (MAD = 0.3 kcal/mol). The considerably more expensive B2PLYP double hybrid^{33,34} does not perform well without the D correction (MAD = 0.6 kcal/mol). However, the B2PLYP-D results show the best agreement with the reference (MAD = 0.2 kcal/mol).¹⁵ One major drawback of B2PLYP is its $O(N)^5$ scaling of computer time with the size N . As this approach uses not only the occupied orbitals but also the unoccupied orbitals, it goes beyond the fourth-rung hyper-GGA approximations, and it can be called a fifth-rung DFT approximation.³⁵ Observe also that in

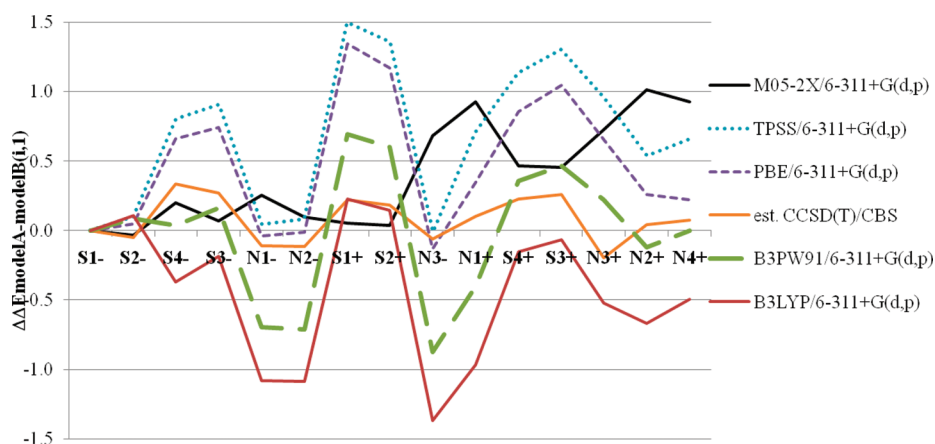


Figure 10. The AnGol15 $\Delta\Delta E_{\text{modelA-modelB}}(i,1)$ relative energy differences (kcal/mol) for modelA = M05-2X/6-311+G(d,p), B3LYP/6-311+G(d,p), B3PW91/6-311+G(d,p), estimated CCSD(T)/CBS, PBE/6-311+G(d,p), and TPSS/6-311+G(d,p) and for modelB = MP2/CBS[4, 5]. All the DFT results are from ref 1 and the estimated CCSD(T)/CBS results are from ref 15. All calculations use the B3LYP/6-31+G(d,p) geometries.

contrast with standard DFT functionals, B2PLYP diverges for bulk metals.

For the ALL15 test set in all the DFT and HF calculations we use the 6-311+G(d,p) basis set and the optimized B3LYP/6-31G(d,p) geometries.¹ The two worst performers are the LSDA and the HF methods. While LSDA systematically overestimates the conformational energy differences (MD = 1.8 and MAD = 1.8 kcal/mol), the HF shows the opposite error (MD = -1.4 and MAD = 1.4 kcal/mol). TPSS, PBE, and M05-2X show improvements compared to the LSDA but conserve some of the LSDA overestimation error (MD = 0.3–0.5 kcal/mol, MAD < 0.6 kcal/mol). The 25% of global mixing of exact exchange decreases the overestimation, and the PBE0 results are improved (MD = 0.2 kcal/mol, RRD = 0.6 kcal/mol). This is a very promising result compared to the other functionals. The remaining small positive deviations might be further reduced by a recently proposed 32% mixing of the exact exchange.³⁶ The best performance is shown by the B3PW91 functional (MAD = 0.16 kcal/mol). The B3LYP functional performs only slightly worse (MAD = 0.2 kcal/mol), but the range of the relative difference is large (RRD = 0.9 kcal/mol). This shows that the B3LYP results will not improve if a different reference conformer is chosen, while the global hybrid of PBE results can be improved further.

Our results show that the nonempirical functionals constructed on the basis of constraint satisfaction, such as PBE and TPSS, perform quite well, and they might outperform the empirical B3LYP or M05-2X functionals. The original B3PW91 functional gives consistently better results than B3LYP. The relative energies given by the best DFT methods are quite close to the reference values with mean absolute deviations around 0.1–0.2 kcal/mol. Thus for correct evaluation of such functionals, accurate reference values are necessary. We plan to test PBEh,³⁶ PBEsol,³⁷ revTPSS,³⁸ M06-2X,³⁹ and M08-HX⁴⁰ functionals in the near future. The benefits of the a posteriori dispersion correction^{32,41} will also be studied.

6. CONCLUSIONS

We have performed a series canonical MP2 complete basis set extrapolations using correlation consistent basis sets up to the cardinal number 5 for the AnGol15, GLC4, and ALL15 monosaccharide test sets. Accurate MP2/CBS[4, 5] reference energies

were obtained, and CCSD(T) corrections were also considered for the relative energies of the AnGol15 and GLC4 test sets. Good agreement was observed with the earlier CCSD(T)/CBS results,¹⁵ and our results are converged for relative energies within 0.1 kcal/mol. This accuracy is necessary for the correct evaluation of DFT methods, as these methods might reach 0.1–0.2 kcal/mol average accuracy. The accuracy of the MP2/aV3Z(-df) results (0.2–0.5 kcal/mol) is generally not enough for correct evaluations. We have observed that the less expensive local MP2 methods do not give consistent results with the 0.03 domain selection criterion for the AnGol15 test set. Further LMP2 studies are necessary to resolve this problem.

Detailed analysis of the convergence of the HF and the MP2 correlation energies with respect to the cardinal number of the basis set revealed that for several conformer pairs little computational effort yields well-converged relative energies. We introduced three criteria that monitor the convergence of the relative energies with respect to the cardinal number. By applying two of these criteria we were able to select conformer pairs for which even HF/V3Z relative energies are well-converged. Fulfillment of the third criterion makes the MP2/V4Z energy well-converged.

The most stable conformers of the AnGol15, GLC4, and ALL15 conformational space are stabilized by the cooperative intramolecular O–H···O interactions and other (e.g., anomeric) electron correlation effects. The HF method misses these effects, seriously underestimating the stability of these conformers compared to the other conformers having weaker stabilizing correlation effects. This leads to too-small relative energies and negative mean deviation from the reference values. Our previous studies showed that LSDA strongly overestimates these stabilization effects, leading to too-large relative energies that show positive mean deviation.¹

The GGAs and meta-GGAs improve on LSDA, but the relative conformational energies might remain statistically too large (mean deviation is smaller but positive).¹ For such GGA functionals an empirical dispersion correction does not improve the results. Global mixing of the GGAs with the exact exchange might further improve the results by reducing overestimation of the relative energies, thus shifting mean deviation close to zero. A sufficiently large exact exchange fraction might even lead to underestimation of the correlation effects. Such global hybrid

functionals can be efficiently corrected by simple and quick empirical dispersion correction.

For the AnGol15 test set, the ordering of the functionals compared to the CCSD(T)/CBS results is, from best to worse: B3PW91, PBE = M05-2X, and TPSS = B3LYP (MAD = 0.3, 0.4, 0.4, 0.6, and 0.6 kcal/mol, respectively).

For the GLC4 test set, the B3PW91 and PBE0 functionals perform equally well (MAD = 0.3 kcal/mol), while B3LYP is somewhat worse (MAD = 0.5 kcal/mol). The M05-2X performs the worst (MAD = 0.8 kcal/mol).

For the ALL15 test set, the B3PW91 performs the best (MAD = 0.16 kcal/mol), followed by PBE0 and a slightly worse B3LYP (MAD = 0.2 kcal/mol). The M05-2X performs somewhat worse (MAD = 0.5 kcal/mol). These results suggest that a new PBE hybrid with 32% weight of the exact exchange might perform better and might be efficiently corrected in the DFT-D or dD10 framework.

These new CCSD(T) and MP2/CBS reference energies modify our previous conclusions.¹ The overall best performer is B3PW91, closely followed by PBE0. Due to the errors in the MP2/aV3Z(-df) energies used as reference in our previous study the good performance of B3PW91 was not this clear, and the performance of the M05-2X was judged considerably better than here. In agreement with our earlier conclusion the B3LYP functional is not the best choice, but the B3LYP-D performs considerably better. In other areas of molecular chemistry the M05-2X and B3LYP perform better than here. The monosaccharide test sets do not sample strong steric interactions or significant dispersion attractions; however, they sample elaborate systems of intramolecular interactions. The correct description of these conformational spaces is required from good model chemistry for biomolecules. The reference energies published here are suitable to evaluate the performance of future KS-DFT and -D or -dD10 corrected functionals with a sufficient accuracy.

■ ASSOCIATED CONTENT

S Supporting Information. The schematic representation of the conformational spaces; the HF, MP2 and LMP2 energies for the AnGol15 and ALL15 test sets. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail csonkagi@gmail.com.

■ ACKNOWLEDGMENT

This work connected to the scientific program of the “Development of quality-oriented and harmonized R+D+I strategy and functional model at BME” project, supported by the New Hungary Development Plan (Project ID: TAMOP-4.2.1/B-09/1/KMR- 2010-0002).

■ REFERENCES

- (1) Csonka, G. I.; French, A. D.; Johnson, G. P.; Stortz, C. A. *J. Chem. Theor. Comput.* **2009**, *5*, 679–692.
- (2) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- (3) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.

- (4) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.
- (5) Ernzerhof, M.; Scuseria, G. E. *J. Chem. Phys.* **1999**, *110*, 5029–5036.
- (6) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623–11627.
- (7) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Kackson, K. A.; Pederson, M. A.; Singh, D. J.; Fiolhais, C. *Phys. Rev. B* **1992**, *46*, 6671–6687.
- (8) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (9) Zhao, Y.; Truhlar, D. G. *Org. Lett.* **2006**, *8*, 5753–5755.
- (10) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *2*, 364–382.
- (11) Kaminsky, J.; Mata, R. A.; Werner, H. J.; Jensen, F. *Mol. Phys.* **2008**, *106*, 1899–1906.
- (12) Pulay, P.; Saebo, S. *Theor. Chim. Acta* **1986**, *69*, 357–368.
- (13) Murphy, R. B.; Beachy, M. D.; Friesner, R. A.; Ringnalda, M. N. *J. Chem. Phys.* **1995**, *103*, 1481–1490.
- (14) Mata, R. A.; Werner, H. J. *J. Chem. Phys.* **2006**, *125*, 184110.
- (15) Goerigk, L.; Grimme, S. *J. Chem. Theor. Comput.* **2010**, *5*, 107–126.
- (16) Werner, H.-J.; Knowles, P. J.; Lindh R. et al. , *Molpro*, version 2006.4; University College Cardiff Consultants Limited: Wales, U.K., 2007; <http://www.molpro.net>. Accessed December 30, 2010.
- (17) Karton, A.; Martin, J. M. L. *Theor. Chem. Acc.* **2006**, *115*, 330–333.
- (18) Jensen, F. *Theor. Chem. Acc.* **2005**, *113*, 187.
- (19) Dunning, T. H.; Peterson, K. A.; Wilson, A. K. *J. Chem. Phys.* **2001**, *114*, 9244.
- (20) Martin, J. M. L. *J. Chem. Phys.* **1998**, *108*, 2791.
- (21) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *J. Chem. Phys.* **1997**, *106*, 9639.
- (22) Kutzelnigg, W.; Morgan, J. D., III *J. Chem. Phys.* **1992**, *96*, 4484.
- (23) Klopper, W. *J. Chem. Phys.* **2001**, *115*, 761.
- (24) Heckert, M.; Kállay, M.; Gauss, J. *Mol. Phys.* **2005**, *103*, 2109.
- (25) Heckert, M.; Kállay, M.; Tew, D. P.; Klopper, W.; Gauss, J. *J. Chem. Phys.* **2006**, *125*, 044108.
- (26) Navarro, D. A.; Stortz, C. A. *Carbohydr. Res.* **2008**, *343*, 2292–2298.
- (27) Cremer, P.; Pople, J. A. *J. Am. Chem. Soc.* **1975**, *97*, 1354.
- (28) *Jaguar 6.0*, release 107; Schrödinger, LLC: Portland, OR, 2005.
- (29) Csonka, G. I. *J. Mol. Struct. (Theochem)* **2002**, *584*, 1–4.
- (30) Csonka, G. I.; Elias, K.; Csizmadia, I. G. *Chem. Phys. Lett.* **1996**, *257*, 49–60.
- (31) Schnupf, U.; Willett, J. L.; Bosma, W. B.; Momany, F. A. *Carbohydr. Res.* **2007**, *342*, 196–216.
- (32) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787–1799.
- (33) Grimme, S. *J. Chem. Phys.* **2006**, *124*, 034108.
- (34) Schwabe, T.; Grimme, S. *Phys. Chem. Chem. Phys.* **2007**, *9*, 3397–3406.
- (35) Perdew, J. P.; Ruzsinszky, A.; Tao, J.; Staroverov, V. N.; Scuseria, G. E.; Csonka, G. I. *J. Chem. Phys.* **2005**, *123*, 62201.
- (36) Csonka, G. I.; Perdew, J. P.; Ruzsinszky, A. *J. Chem. Theor. Comput.* **2010**, *6*, 3688–3703.
- (37) Perdew, J. P.; Ruzsinszky, A.; Csonka, G. I.; Vydrov, O. A.; Scuseria, G. E.; Constantin, L. A.; Zhou, X.; Burke, K. *Phys. Rev. Lett.* **2008**, *100*, 136406.
- (38) Perdew, J. P.; Ruzsinszky, A.; Csonka, G. I.; Constantin, L. A.; Sun, J. *Phys. Rev. Lett.* **2009**, *103*, 026403.
- (39) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *2*, 364.
- (40) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 1849.
- (41) Steinmann, S. N.; Csonka, G. I.; Corminboeuf, C. *J. Chem. Theor. Comput.* **2009**, *5*, 2950 and references cited therein.

Parallel Implementation of the Four-Component Relativistic Quasidegenerate Perturbation Theory with General Multiconfigurational Reference Functions

Ryo Ebisuzaki,[†] Yoshihiro Watanabe,[†] Yukio Kawashima,^{†,‡} and Haruyuki Nakano^{*,†}

[†]Department of Chemistry, Graduate School of Sciences, Kyushu University, Fukuoka 812-8581, Japan

[‡]Institute of Advanced Research, Kyushu University, Fukuoka 812-8581, Japan

ABSTRACT: A new, efficient parallel algorithm for four-component relativistic generalized multiconfigurational quasidegenerate perturbation theory (GMC-QDPT) introducing Kramers symmetry is implemented. Because it utilizes the independence of the terms in the matrix element computations, this algorithm both speeds up the calculation and reduces the computational resources required for each node. In addition, the amount of memory for two-electron integrals is reduced to three-eighths by Kramers restriction. The algorithm is applied to the d–d excitation energies of the platinum halide complexes, $[\text{PtCl}_4]^{2-}$, $[\text{PtBr}_4]^{2-}$, and $[\text{PtCl}_6]^{2-}$ and to the 6p–7s and 6p–7p excitation energies of the radon atom. It is shown to provide high parallelization efficiency and accurate excitation energies that agree well with experimental data.

1. INTRODUCTION

In electronic structure calculations for systems that contain heavy elements, the inclusion of both the electron correlation and relativistic effects is essential for high accuracy. Thus, many electron correlation methods have been extended to four-component relativistic versions, including Møller–Plesset perturbation,^{1–5} configuration interaction (CI),^{6,7} and coupled cluster methods,^{8,9} which are based on the Dirac–Hartree–Fock (DHF) reference wave function. In addition, four-component multireference (MR) CI^{10–12} and coupled cluster¹³ methods that provide highly accurate molecular electronic structures are also now available. However, the four-component MR methods are computationally expensive, and their applications are therefore limited to small molecules.

Among the MR methods, multireference perturbation theory (MRPT) is efficient and accurate and thus is regarded as a practical tool that takes account of both static and dynamic electron correlations. Recently, we developed a MRPT using the general multiconfigurational functions called GMC-QDPT or GMC-PT.¹⁴ GMC-QDPT is applicable to any multiconfigurational reference wave functions and allows the use of only necessary configurations according to the character of the system of interest. Because of its generality and flexibility, GMC-QDPT enjoys both high computational accuracy and efficiency. However, it is still not easy to calculate heavy atom compounds with many electrons efficiently using relativistic GMC-QDPT.¹⁵ Therefore, there is a need for computational schemes for relativistic GMC-QDPT that are more efficient than the existing one.

Previously, we proposed a new efficient calculation scheme for the effective Hamiltonian based on matrix elements between the reference and ionized functions and succeeded in reducing computation time relative to the previous scheme.¹⁶ In the present work, we developed a parallel

algorithm based on Kramers symmetry. For the nonrelativistic MC-QDPT, Umeda and co-workers have presented a parallel algorithm.¹⁷ This algorithm was based on the previous scheme of ours using diagrams and thus has a disadvantage for the computation of one virtual terms (the terms involving two-electron integrals with one virtual and three occupied orbital labels) of the effective Hamiltonian.¹⁶ In addition, we applied the new code accelerated by our parallel algorithm to the calculations for the d–d excitation energies of platinum halide complexes ($[\text{PtCl}_4]^{2-}$, $[\text{PtBr}_4]^{2-}$, and $[\text{PtCl}_6]^{2-}$) and the 6p–7s/6p–7p excitation energies of the radon atom.

In Section 2, we review GMC-QDPT and describe our new parallel algorithm and its implementation. We report the efficiency of our new scheme and the accuracy of the excitation energy calculation results in Section 3.

2. METHODS

2.1. Brief Review of GMC-QDPT. Before describing the new computational scheme, we briefly review the GMC-QDPT for readers' convenience.

Let $|\mu\rangle$ and $E_\mu^{(0)}$ be reference (zeroth-order) wave functions and their zeroth-order energies:

$$|\mu\rangle = \sum_{A \in \text{GCS}} C_A^\mu |A\rangle, \quad E_\mu^{(0)} = \sum_P \langle \mu | a_P^\dagger a_P | \mu \rangle \varepsilon_P \quad (1)$$

respectively. Here, the reference wave functions $|\mu\rangle$ are expanded by the determinants $|A\rangle$ in a general configuration space (GCS); a_P^\dagger and a_P are the creation and annihilation

Received: January 7, 2011

Published: March 18, 2011

operators, respectively, for an electron in spinor P ; and ε_P are spinor energies. The effective Hamiltonian up to the second order of GMC-QDPT^{14,15} is given as

$$(H_{\text{eff}}^{(0-2)})_{\mu\nu} = \langle \mu | H | \nu \rangle - \frac{1}{2} \left\{ \sum_{I \notin \text{GCS}} \frac{\langle \mu | H | I \rangle \langle I | H | \nu \rangle}{E_\nu^{(0)} - E_I^{(0)}} + \sum_{I \notin \text{GCS}} \frac{\langle \mu | H | I \rangle \langle I | H | \nu \rangle}{E_\mu^{(0)} - E_I^{(0)}} \right\} \quad (2)$$

where $|I\rangle$ is the determinant outside the GCS.

We define the corresponding complete active space (CCAS) as the minimal CAS including the reference GCS. The summation over $|I\rangle$ can be divided into a summation outside the CCAS and a summation inside the CCAS but outside the GCS. Thus, the summation enclosed in curly brackets in eq 2 can be written as

$$(H_{\text{eff}}^{(2)})_{\mu\nu} = \sum_{I \notin \text{CCAS}} \frac{\langle \mu | H | I \rangle \langle I | H | \nu \rangle}{E_\nu^{(0)} - E_I^{(0)}} + \sum_{I \in \text{CCAS} \wedge I \notin \text{GCS}} \frac{\langle \mu | H | I \rangle \langle I | H | \nu \rangle}{E_\nu^{(0)} - E_I^{(0)}} \quad (3)$$

We call the terms in the first summation “the external terms” and the terms in the second summation “the internal terms”.¹⁴

The internal terms are computed through matrix operations for the Hamiltonian matrix. The length of the internal term summation is generally much less than that of the external term summation. Hence, the computational time for the internal terms is small compared with that for the external terms.

The external terms are computed using a matrix element scheme.¹⁶ The intermediate states $|I\rangle$ can be written as a product of determinants comprising M active spinors $|B^M\rangle$ and ionization/excitation operators:

$$E_{QS\dots U}^{PR\dots T} = a_p^\dagger a_r^\dagger \dots a_T^\dagger a_U \dots a_S a_Q \quad (4)$$

(abbreviated as E_X). Note that the numbers of creation and annihilation operators are not necessarily equal.

Hence, the summation of the external terms is expressed as

$$(H_{\text{external}}^{(2)})_{\mu\nu} = \sum_X \sum_{B^M} \frac{\langle \mu | H E_X | B^M \rangle \langle B^M | E_X^\dagger H | \nu \rangle}{E_\nu^{(0)} - E_{XB^M}^{(0)}} \quad (5)$$

where $E_{XB^M}^{(0)}$ are zeroth-order energies of $E_X|B^M\rangle$. In this scheme, we calculate eq 5 as the summation of the product of each matrix element $\langle \mu | H E_X | B^M \rangle$. There are five combinations of E_X and $|B^M\rangle$ that give nonzero $\langle \mu | H E_X | B^M \rangle$. Each matrix element can be simplified through normal ordering of creation and annihilation operators. These can be divided into two categories based on the number of virtual spinor labels. We define the terms with one or no virtual spinor label as 1-virtual terms and other terms with two virtual spinor labels as 2-virtual terms. For example, one of the 1-virtual

terms is

$$\begin{aligned} \langle \mu | H E^E | B^{N-1} \rangle &= \sum_P f_{PE}^{\text{core}} \langle \mu | E^P | B^{N-1} \rangle \\ &+ \frac{1}{2} \sum_{PRS} (PE \parallel RS) \langle \mu | E_S^{PR} | B^{N-1} \rangle \\ &= \sum_P \sum_A f_{PE}^{\text{core}} C_A^{\mu*} \langle A | E^P | B^{N-1} \rangle \\ &+ \frac{1}{2} \sum_{PRS} \sum_A (PE \parallel RS) C_A^{\mu*} \langle A | E_S^{PR} | B^{N-1} \rangle \quad (6) \end{aligned}$$

and one of the 2-virtual terms is

$$\begin{aligned} \langle \mu | H E^{EF} | B^{N-2} \rangle &= \frac{1}{2} \sum_{PR} (PE \parallel RF) \langle \mu | E^{PR} | B^{N-2} \rangle \\ &= \frac{1}{2} \sum_{PR} \sum_A (PE \parallel RF) C_A^{\mu*} \langle A | E^{PR} | B^{N-2} \rangle \quad (7) \end{aligned}$$

where E and F are virtual spinor labels; I and J are core spinor labels; P , Q , and R are active spinor labels; N is the number of active electrons; f_{PE}^{core} are the core Fock matrix elements; $(PQ \parallel RS)$ are the antisymmetrized two-electron integrals; and $\langle A | E_X | B^M \rangle$ are the coupling coefficients (CCs).

2.2. Parallel Algorithm. A feature of the perturbation method at the second order is that the energy or effective Hamiltonian is expressed as a summation of many independent terms. Neither diagonalization of large matrices nor solution of large linear equations is necessary. Specifically, the terms for determinant B^M and ionization operator X in eq 5 are independent of each other. In addition, the matrix elements used to obtain the terms in eq 5 are also simple sums of the product of molecular integrals, CI coefficients, and CCs, as seen in eqs 6 and 7.

The most straightforward method of parallelization is to use the independency of the terms for the determinants B^M in eq 5. The summation for B^M can be computed easily in parallel after being divided and assigned to computational nodes. In fact, the speedup by parallel computing was almost linear with the number of computational nodes in our preliminary calculations. However, in this parallelization, all molecular integrals are required for each node, and thus no scalability is gained for integral storage, which is a real problem in practice. Therefore, parallelization was done utilizing the independency of the terms in the matrix element computations of eqs 5 and 6.

In the serial algorithm, the Hamiltonian matrix elements and their associated effective Hamiltonian matrix elements are calculated according to Schemes 1 and 2 for $\langle \mu | H E^E | B^{N-1} \rangle$ and $\langle \mu | H E^{EF} | B^{N-2} \rangle$, respectively. The algorithm is coupling coefficient driven. For each ionized determinant $|B^M\rangle$, all nonzero coupling coefficients $\langle A | E_X | B^M \rangle$ ($= +1$ or -1) are calculated first. Then, being operated by an ionization operator E_X , intermediate determinants $|I\rangle = E_X | B^M \rangle$ are made, and the matrix elements $\langle \mu | H E_X | B^M \rangle$ are computed. Finally, the effective Hamiltonian elements are computed.

The steps specific on the parallel algorithm are emphasized in bold in Schemes 1 and 2. Consider Scheme 2 as an example to explain the parallel algorithm. Usually, the computational cost of the coupling coefficients $\langle A | E_X | B^M \rangle$ is much smaller than that of the matrix elements $\langle \mu | H E_X | B^M \rangle$. For each B^{N-2} , the operation count for $\langle A | H E^{PR} | B^{N-2} \rangle$ is $O(m_{\text{act}}^2)$, where m_{act} is the number of active spinors, while the operation count for $\langle \mu | H E^{EF} | B^{N-2} \rangle$ is $O(m_{\text{act}}^2 m_{\text{vir}}^2)$, where m_{vir} is the number of virtual spinors. Thus, in

Scheme 1. Loop Structure for the Effective Hamiltonian Contributed From 1-Virtual Integrals^a

- Loop over ionized determinants $|B^{N-1}\rangle$
- Calculate all nonzero coupling coefficients (CCs) $\langle A|E^P|B^{N-1}\rangle$ for $|B^{N-1}\rangle$
 - Calculate all nonzero CCs $\langle A|E_S^{PR}|B^{N-1}\rangle$ ($P < R$) for $|B^{N-1}\rangle$
 - Distribute E to computational nodes**
 - On each node: Loop over E assigned to the node**
 - Clear $h(\dots)$
 - Loop over nonzero CCs $\langle A|E^P|B^{N-1}\rangle$
Loop over reference states μ
If $\langle A|E^P|B^{N-1}\rangle = \pm 1$, $h(\mu) = h(\mu) \pm f_{PE}^{\text{core}} C_A^{\mu*}$
 - Loop over nonzero CCs $\langle A|E_S^{PR}|B^{N-1}\rangle$
Loop over reference states μ
If $\langle A|E_S^{PR}|B^{N-1}\rangle = \pm 1$, $h(\mu) = h(\mu) \pm (PE||RS) C_A^{\mu*}$
 - Loop over reference states μ and ν
 $H_{\text{eff}}(\mu, \nu) = H_{\text{eff}}(\mu, \nu) + h(\mu) \cdot h(\nu)^* / (E_\nu^{(0)} - E_{B^{N-1}}^{(0)} - \epsilon_E)$
 - Calculate global sum of $H_{\text{eff}}(\dots, \dots)$**

^a The steps required in parallel algorithm are shown in bold.

Scheme 2. Loop Structure for the Effective Hamiltonian Contributed From 2-Virtual Integrals^a

- Loop over ionized determinants $|B^{N-2}\rangle$
- Calculate all nonzero coupling coefficients (CCs) $\langle A|E^{PR}|B^{N-2}\rangle$ ($P < R$) for $|B^{N-2}\rangle$
 - Distribute E, F pairs to computational nodes ($E < F$)**
 - On each node: Loop over E and F assigned to the node**
 - Clear $h(\dots)$
 - Loop over nonzero CCs $\langle A|E^{PR}|B^{N-2}\rangle$
Loop over reference states μ
If $\langle A|E^{PR}|B^{N-2}\rangle = \pm 1$, $h(\mu) = h(\mu) \pm (PE||RF) C_A^{\mu*}$
 - Loop over reference states μ and ν
 $H_{\text{eff}}(\mu, \nu) = H_{\text{eff}}(\mu, \nu) + h(\mu) \cdot h(\nu)^* / (E_\nu^{(0)} - E_{B^{N-2}}^{(0)} - \epsilon_E - \epsilon_F)$
 - Calculate global sum of $H_{\text{eff}}(\dots, \dots)$**

^a The steps required in parallel algorithm are shown in bold.

the parallel algorithm, the coupling coefficients are computed first in each computational node. Next, the E, F pairs are divided and assigned to computational nodes. (In other words, the ionization/excitation operators $E_X = E^{EF}$ are distributed.) Then, part of the effective Hamiltonian is calculated on each node for the assigned E, F pairs. Finally, the effective Hamiltonians on respective nodes are collected and summed to obtain the total effective Hamiltonian. The parallel algorithms for the other terms are similar to Schemes 1 and 2.

The speedup by the parallel algorithm is roughly estimated as

$$S_{\text{estimated}}(n) = \frac{T_{\text{EH}} + T_{\text{CC}}}{T_{\text{EH}}/n + T_{\text{CC}}} \quad (8)$$

where n is the number of computational nodes, and T_{EH} and T_{CC} are the computational times of the serial algorithm for

the perturbation summation and CCs, respectively. As long as T_{CC} is negligible compared with T_{EH}/n (namely $T_{\text{EH}}/n \gg T_{\text{CC}}$), parallelization is expected to speed up the calculation by a factor of n .

The integral storage is also reduced by a factor n by parallelization. For the 1-virtual terms in Scheme 1, label E and associated integrals $(PE||RS)$ is distributed to each computational node. Hence the integral storage on a node is reduced to $(\lceil (m_{\text{vir}} - 1)/n \rceil + 1) m_{\text{vir}}^{-1}$ of the total, where the Gauss bracket $\lceil X \rceil$ denotes the largest integer less than or equal to X . For the 2-virtual terms in Scheme 2, since the pair of labels (E, F) and associated integrals $(PE||RF)$ ($E < F$) is distributed, the integral storage on a node is reduced to $(\lceil (m_{\text{vir}}(m_{\text{vir}} - 1)/2 - 1)/n \rceil + 1) (m_{\text{vir}}(m_{\text{vir}} - 1)/2)^{-1}$ of the total. These values approach $1/n$ if m_{vir} is sufficiently large compared with n .

2.3. Kramers Restriction (Kramers-Restricted GMC-QDPT Formulas). The original GMC-QDPT was expressed in the

Kramers-unrestricted form as in eqs 6 and 7 to allow for the external magnetic field when needed. However, Kramers-restricted formulas allow us to use only unique integrals in the absence of an external magnetic field, such as the implementation of Yanai et al.¹⁸ in a 4-spinor molecular Dirac–Fock SCF method. In our implementation, we employ Kramers restriction to save memory resources for GMC-QDPT calculation.

The time-reversed function $\bar{\phi} = \phi(-t)$ is written through the time reversal operator \hat{K} ; as

$$\hat{K}\phi = \bar{\phi}, \quad \hat{K}\bar{\phi} = -\phi \quad (9)$$

Using the relation

$$\hat{K}(pq||rs) = (-1)^N (\hat{K}p\hat{K}q||\hat{K}r\hat{K}s)^* \quad (10)$$

where $\bar{K}p$ is a short expression for $\hat{K}\phi_p$, and N is the number of barred spinors in $\{\phi_p, \phi_q, \phi_r, \phi_s\}$ and the label symmetry:

$$(pq||rs) = -(rq||ps) = -(ps||rq) = (rs||pq) \quad (11)$$

we can reduce the 16 ($= 2^4$) kinds of 1- and 2-virtual integrals to 6:

$$\begin{aligned} (pe||rs) &= (\bar{p}\bar{e}||\bar{r}\bar{s})^*, & (\bar{p}\bar{e}||\bar{r}\bar{s}) &= -(p\bar{e}||rs)^* \\ (pe||r\bar{s}) &= -(\bar{p}\bar{e}||\bar{r}s)^*, & (\bar{p}\bar{e}||\bar{r}s) &= (p\bar{e}||r\bar{s})^* \\ (pe||\bar{r}s) &= -(\bar{p}\bar{e}||r\bar{s})^* = -(\bar{r}\bar{e}||ps) = (r\bar{e}||\bar{p}\bar{s})^* \\ (\bar{p}\bar{e}||r\bar{s}) &= (p\bar{e}||\bar{r}\bar{s})^* = -(r\bar{e}||\bar{p}\bar{s}) = -(\bar{r}\bar{e}||ps)^* \end{aligned} \quad (12)$$

for 1-virtual integrals, and

$$\begin{aligned} (pe||rf) &= (\bar{p}\bar{e}||\bar{r}\bar{f})^*, & (\bar{p}\bar{e}||\bar{r}\bar{f}) &= -(p\bar{e}||rf)^* \\ (pe||r\bar{f}) &= -(\bar{p}\bar{e}||\bar{r}f)^*, & (\bar{p}\bar{e}||\bar{r}f) &= (p\bar{e}||r\bar{f})^* \\ (\bar{p}\bar{e}||r\bar{f}) &= -(\bar{p}\bar{e}||\bar{r}\bar{f})^* = -(r\bar{e}||\bar{p}\bar{f}) = (\bar{r}\bar{e}||p\bar{f})^* \\ (pe||\bar{r}\bar{f}) &= (\bar{p}\bar{e}||rf)^* = -(\bar{r}\bar{e}||p\bar{f}) = -(r\bar{e}||\bar{p}\bar{f})^* \end{aligned} \quad (13)$$

for 2-virtual integrals.

By reducing the integrals, we can obtain Kramers-restricted formulas for the matrix elements as eqs 6 and 7. For example, $\langle\mu|HE^{EF}|B^{N-2}\rangle$ are rewritten as

$$\begin{aligned} \langle\mu|HE^{ef}|B^{N-2}\rangle &= \frac{1}{2} \sum_{pr} \sum_A \{ (pe||rf) C_A^{\mu*} \langle A|E^{pr}|B^{N-2}\rangle \\ &+ 2(\bar{p}\bar{e}||rf) C_A^{\mu*} \langle A|E^{\bar{p}\bar{r}}|B^{N-2}\rangle + (\bar{p}\bar{e}||\bar{r}\bar{f}) C_A^{\mu*} \langle A|E^{\bar{p}\bar{r}}|B^{N-2}\rangle \} \end{aligned} \quad (14)$$

$$\begin{aligned} \langle\mu|HE^{\bar{e}\bar{f}}|B^{N-2}\rangle &= \frac{1}{2} \sum_{pr} \sum_A \{ (pe||rf)^* C_A^{\mu*} \langle A|E^{\bar{p}\bar{r}}|B^{N-2}\rangle \\ &+ 2(\bar{p}\bar{e}||r\bar{f})^* C_A^{\mu*} \langle A|E^{p\bar{r}}|B^{N-2}\rangle + (\bar{p}\bar{e}||\bar{r}\bar{f})^* C_A^{\mu*} \langle A|E^{p\bar{r}}|B^{N-2}\rangle \} \end{aligned} \quad (15)$$

$$\begin{aligned} \langle\mu|HE^{\bar{e}\bar{f}}|B^{N-2}\rangle &= \frac{1}{2} \sum_{pr} \sum_A \{ (\bar{p}\bar{e}||\bar{r}\bar{f}) C_A^{\mu*} \langle A|E^{\bar{p}\bar{r}}|B^{N-2}\rangle \\ &+ 2(pe||\bar{r}\bar{f}) C_A^{\mu*} \langle A|E^{p\bar{r}}|B^{N-2}\rangle + (pe||r\bar{f}) C_A^{\mu*} \langle A|E^{p\bar{r}}|B^{N-2}\rangle \} \end{aligned} \quad (16)$$

(Type $\langle\mu|HE^{\bar{e}\bar{f}}|B^{N-2}\rangle$ do not appear because of the restriction $E < F$.) These formulas are actually used in the program.

In contrast to Kramers-unrestricted formulas, the length of the summation for Kramers-pair labels p and r in eqs 14–16 is one-fourth of that for spinor labels P and R in eq 7, whereas the number of terms in the summation has increased from one to three. Therefore, the operation count is hardly reduced. On the other hand, however, we can reduce the amount of memory used to store integrals to about three-eighths ($= 6/16$).

3. RESULTS AND DISCUSSION

We applied the present computational scheme to some molecular systems and measured CPU time (on 3.0 GHz Pentium D 930 processors) to evaluate its performance. We calculated the d–d excitation energies for three platinum halide anions (d⁸-complex [PtCl₄]²⁻, [PtBr₄]²⁻, and d⁶-complex [PtCl₆]²⁻), and the 6p–7s and 6p–7p excitation energies of the Rn atom.

The spinors used in the perturbation calculations were determined using the four-component DHF method¹⁹ using the REL4D program²⁰ of UTChem.²¹ The basis set proposed by Koga, Tatewaki, and Matsuoka (KTM)²² was employed for the platinum halide complex calculations. For Cl, a d basis spinor (exponent 0.514) was added as a polarization function. For Rn, Dyal's triple- ζ ²³ basis set, which includes up to g-type polarization function, augmented by single s, p, and d diffuse functions in an even-tempered manner was employed. The molecular structures for the platinum halide complexes were taken from experimental data.^{24,25} The zeroth-order wave functions were set according to the scheme for nonrelativistic multireference multi state perturbation theory implemented in Firefly.²⁶

3.1. d–d Excitation Energies of Platinum Tetrachloride Dianion [PtCl₄]²⁻. First, we calculated the d–d excitation energies of [PtCl₄]²⁻. The target states were the 12 lowest excited states resulting from d–d single excitations as well as the ground state. The reference CI space was of a multireference singles (MRS) type constructed from 20 electrons and 26 spinors, which include 5d components of the platinum atom largely and therefore necessary to describe the considered excitation states. The determinants that spanned the reference CI space were generated from 41 parent configurations (the DHF configuration and the singly excited configurations constructed from the highest 20 occupied and the lowest 2 unoccupied spinors) and selected according to their weights in the reference functions. The determinants whose weights in the reference wave functions were greater than 10⁻⁸ (i.e., $|C_i| > 10^{-4}$) were selected. The lowest 108 spinors were frozen in the perturbation calculations. Compared with previous papers,¹⁶ we used larger basis sets that were specifically designed for four-component relativistic calculations.

First, let us discuss the accuracy of the GMC-QDPT results.

The computed excitation energies are summarized in Table 1, together with experimental data from Patterson²⁷ and the two-component time-dependent density functional theory (TDDFT) results of Wang and Ziegler²⁸ for comparison. For the states for which experimental data are available (the 2A_{1g}, 1B_{1g}–2B_{1g},

Table 1. d–d Excitation Energies of $[\text{PtCl}_4]^{2-}$ (eV)

state	ref-CI	GMC-QDPT	ref (%) ^a	TDDFT	band	expt
1A _{1g}	2.13	1.98	74.3	2.30		
1A _{2g}	2.20	2.06	74.4	2.34		
1E _g	2.24	2.11	73.6	2.38	1	2.12
1B _{2g}	2.31	2.40	67.7	2.49	2	2.24
1B _{1g}	2.59	2.47	69.8	2.59	3	2.57
2E _g	2.63	2.56	70.8	2.69	3	2.57
2A _{1g}	3.15	2.72	73.7	2.98	4	2.97
3E _g	3.19	2.94	70.1	3.03	4	2.97
2A _{2g}	3.54	3.10	73.9	3.19		
2B _{2g}	3.49	3.38	71.5	3.43	5	3.23
4E _g	3.85	3.54	73.5	3.50	6	3.67
2B _{1g}	3.89	3.82	71.0	3.53	6	3.67

^aThe reference weight of the ground state was 75.5%.

1B_{2g}–2B_{2g}, and 1E_g–4E_g states), the computed values showed good agreement with experimental values. The maximum and average deviations from the experiment were 0.16 and 0.08 eV, respectively.

Table 1 includes the approximate weight that the reference function occupied in the first-order perturbed wave function.¹⁵ GMC-QDPT enables us to include only configurations necessary to construct the reference wave function, which greatly reduces the computational time and resources needed compared with complete active space self-consistent field (CASSCF) reference perturbation theories. However, we must carefully investigate whether we have considered enough configurations in our calculations, because we reduced a large number of configurations to construct our reference wave function. To investigate our selection of the configurations, the weight of the reference function in the first-order perturbed wave function was examined. This weight is a measure of the quality of the reference wave functions; if the weight is large enough, then we have included enough configurations. By comparing the relative weight calculated for different states, we can investigate the balance of the calculation; if the weights for each state have similar values, then we have included enough configurations to describe all states with the same quality. The weights in Table 1 are close to each other. The values are in the range 67.7–74.4% for the excited states and 75.5% for the ground state, which indicates that the calculations were well balanced.

In comparison with the TDDFT results, the GMC-QDPT excitation energies were smaller, especially in the lowest few excited states. TDDFT calculation results showed a tendency to overestimate the excitation energies, whereas GMC-QDPT results agreed well with the experimental data for this system.

Now, let us discuss the efficiency of the parallel algorithm.

Table 2 summarizes the wall computational time and speedup of the present scheme for the $[\text{PtCl}_4]^{2-}$ calculations, and Figure 1 is a plot of speedup ratios. Speedup is defined as $S(n) = T_{\text{seq}}/T(n)$, where T_{seq} is the wall clock time of a sequential execution, and $T(n)$ is the wall clock time of a parallel execution on n nodes.

As seen from Figure 1, the total speedup was approximately proportional to the number of nodes n . The speedup values of calculations for internal, 1-virtual, and 2-virtual terms showed different behaviors. Speedup of the calculation for 2-virtual terms was nearly proportional to the number of nodes; by contrast, the calculation for internal terms showed almost no speedup for an

Table 2. Wall Times and Parallel Efficiency of the $[\text{PtCl}_4]^{2-}$ Calculations^a

no. of nodes	wall time (s)				speedup	parallel efficiency
	internal part	1-virtual part	2-virtual part	total		
1	6.51	137.01	1324.93	1468.44	1.00	1.00
2	5.87	72.34	651.59	729.80	2.01	1.01
3	5.68	48.98	428.97	483.64	3.04	1.01
4	5.62	38.94	313.34	357.89	4.10	1.03
5	5.57	33.63	251.18	290.39	5.06	1.01
6	5.55	29.45	206.05	241.06	6.09	1.02
7	5.53	25.78	173.15	204.46	7.18	1.03
8	5.49	24.08	153.23	182.80	8.03	1.00
9	5.44	22.58	134.63	162.65	9.03	1.00
10	5.42	21.31	119.35	146.08	10.05	1.01
11	5.34	20.18	109.40	134.92	10.88	0.99
12	5.34	19.20	98.17	122.72	11.97	1.00
13	5.29	19.05	89.93	114.27	12.85	0.99
14	5.28	17.55	83.66	106.50	13.79	0.98
15	5.29	17.46	77.50	100.25	14.65	0.98
16	5.28	16.75	70.78	92.81	15.82	0.99
20	5.26	15.67	57.96	78.89	18.61	0.93
24	5.22	14.93	48.63	68.78	21.35	0.89
28	5.25	13.23	41.66	60.14	24.42	0.87
32	5.22	13.21	37.25	55.68	26.37	0.82

^aThe wall times for the CC calculations were 5.08, 7.65, and 0.31 s for the internal, 1-virtual, and 2-virtual parts, respectively.

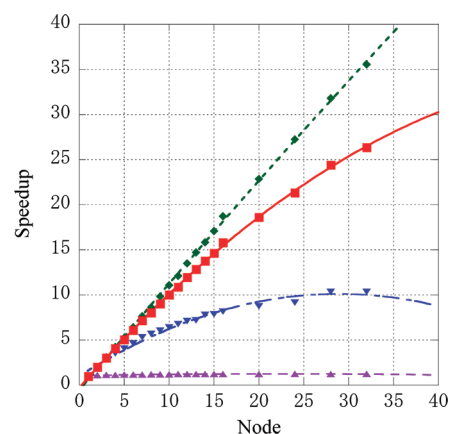


Figure 1. Speedup of the $[\text{PtCl}_4]^{2-}$ calculations. Inverted triangles (∇): 1-virtual part; diamonds (\blacklozenge): 2-virtual part; triangles (\blacktriangle): internal part; and squares (\blacksquare): total.

increase in the number of nodes. The calculation for 1-virtual terms showed intermediate behavior: the speedup increased more slowly as the number of nodes increased. Table 2 also includes parallel efficiency, which is defined by $E(n) = S(n)/n$. The parallel efficiency was high for 2–32 nodes, as implied by the near proportionality of the speedup. The values were greater than 0.80.

The behaviors seen in the calculation of 1- and 2-virtual terms can be explained mainly by the fraction of computational time for the coupling coefficients $\langle A|E_X|B^M \rangle$ occupying by the computation

Table 3. d–d Excitation Energies of $[\text{PtBr}_4]^{2-}$ (eV)

state	ref-CI	GMC-QDPT	ref (%) ^a	TDDFT	band	expt
1A _{1g}	1.94	1.75	76.9	1.93		
1A _{2g}	1.99	1.81	76.9	1.97	1	2.11
1E _g	2.00	1.85	76.9	2.01	1	2.11
1B _{2g}	2.04	2.01	77.4	2.11		
1B _{1g}	2.32	2.13	77.0	2.19		
2E _g	2.36	2.23	77.0	2.29	2	2.37
2A _{1g}	2.98	2.50	76.0	2.66	3	2.81
3E _g	2.98	2.60	76.5	2.67	3	2.81
2A _{2g}	3.29	2.83	76.4	2.82	4	3.02
2B _{1g}	3.23	2.99	76.0	2.72		
4E _g	3.52	3.15	76.4	3.06	5	3.32
2B _{2g}	3.57	3.37	76.2			

^aThe reference weight of the ground state was 78.2%.

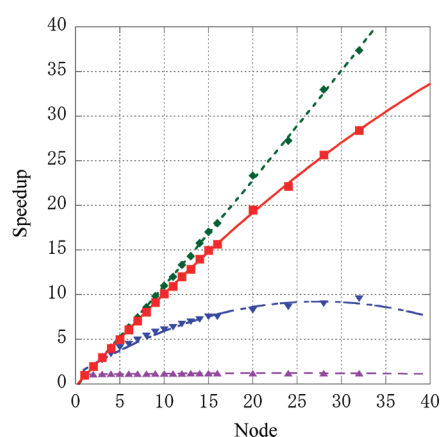


Figure 2. Speedup of the $[\text{PtBr}_4]^{2-}$ calculations. Inverted triangles (\blacktriangledown): 1-virtual part; diamonds (\blacklozenge): 2-virtual part; triangles (\blacktriangle): internal part; and squares (\blacksquare): total.

of each term. In the parallel algorithm explained in Section 2.2, the computational time required for the coupling coefficients was presumed to be small, and thus the coefficients are computed in each node. The speedup can be roughly estimated according to eq 8. For the 2-virtual term, this presumption was satisfactory. The computational time for CCs was only 0.31 s and hence $S(n)$ was close to n . By contrast, for the 1-virtual part, the presumption was not satisfactory. The computational time for CCs was 7.65 s, which was 5.6% of the time required for the 1-virtual term on one node (137.01 s) but 58% on 32 nodes (13.21 s). The computational times of CCs were therefore not negligible in multiple-node calculations, and as a result the $S(n)$ curve behaved as seen in Figure 1.

3.2. d–d Excitation Energies of Platinum Tetrabromide Dianion $[\text{PtBr}_4]^{2-}$. The second example was $[\text{PtBr}_4]^{2-}$. The computational details were similar to those for $[\text{PtCl}_4]^{2-}$. The target excited states of $[\text{PtBr}_4]^{2-}$ were the 12 excited states resulting from d–d single excitations. The reference space was of MRS type constructed from 20 electrons and 28 spinors comprising mostly d spinors of the platinum atom. The determinants that spanned the reference space were prepared in the same manner as in the $[\text{PtCl}_4]^{2-}$ calculation. The lowest 180 spinors were frozen in the perturbation calculations.

The computed excitation energies are summarized in Table 3 together with the experimental data of Kroening and co-workers²⁹

Table 4. d–d Excitation Energies of $[\text{PtCl}_6]^{2-}$ Calculations (eV)

state	ref-CI	GMC-QDPT	ref (%) ^a	TDDFT	expt
1E _g	3.49	2.25	64.8	2.43	
1T _{2g}	3.57	2.31	64.8	2.50	2.23
1T _{1g}	3.62	2.36	65.0	2.49	
1A _{2g}	4.16	2.68	65.0	2.63	
2T _{1g}	4.17	2.76	65.0	2.72	2.64
2E _g	4.42	2.93	64.5	2.74	
2T _{2g}	4.56	3.04	64.4	2.73	
3T _{1g}	4.53	3.05	64.7	2.79	
2A _{2g}	4.66	3.12	64.2	2.88	
3T _{2g}	5.05	3.47	64.5	2.87	3.51

^aThe reference weight of the ground state was 68.5%.

and the TDDFT results.²⁸ For the states for which experimental data are available (2A_{1g}, 1A_{2g}–2A_{2g}, and 2E_g–4E_g states), the computed values were in good agreement with the experimental values, as in the case of $[\text{PtCl}_4]^{2-}$. The maximum and average deviations from the experiment were 0.26 and 0.19 eV, respectively. The GMC-QDPT calculated values were close to the TDDFT values of Wang and Ziegler. Both results were somewhat small compared with the experimental values.

Figure 2 is a plot of speedup ratios. The speedup and parallel efficiency for $[\text{PtBr}_4]^{2-}$ showed similar trends to those found in $[\text{PtCl}_4]^{2-}$. The parallel efficiency was greater than 0.85 for n in the range 2–32.

3.3. d–d Excitation Energies of Platinum Hexachloride Dianion $[\text{PtCl}_6]^{2-}$. The third example was $[\text{PtCl}_6]^{2-}$, which is a d⁶ octahedral complex. The target states were the lowest 10 excited states resulting from d–d single excitations as well as the ground state. The reference space was of MRS type constructed from 16 electrons and 24 spinors mostly comprising d spinors of the platinum atom. The determinants that spanned the reference space were generated from 65 parent configurations (the DHF configuration and the singly excited configurations constructed from the highest 16 occupied and the lowest 4 unoccupied spinors) and then selected according to their weights in the reference functions. The lowest 128 spinors were frozen in the perturbation calculations.

The computed results are summarized in Table 4 together with the experimental^{30,31} data and the TDDFT results.³² State assignment in GMC-QDPT was done in O_h point group symmetry, and calculations were done in D_{4h} group symmetry. For this system, a few experimental numbers (2.23, 2.64, and 3.51 eV) were available, and these were computed to be 2.31, 2.76, and 3.47 eV by GMC-QDPT. Compared with the $[\text{PtCl}_4]^{2-}$ and $[\text{PtBr}_4]^{2-}$ cases, the reference weights were slightly lower because the present calculation involved more correlated electrons. For the same reason, the differences between the reference CI and GMC-QDPT excitation energies were larger: The average difference was 0.18 eV in the $[\text{PtCl}_4]^{2-}$ case compared with 1.43 eV in the $[\text{PtCl}_6]^{2-}$ case.

Figure 3 is a plot of speedup for $[\text{PtCl}_6]^{2-}$ calculations. Unlike the previous cases, speedup of the 1-virtual term was almost proportional to the number of nodes. The total computational time for $[\text{PtCl}_6]^{2-}$ calculation was 7584.54 s and about five times larger than the former two systems. Therefore, the computational time for the CCs (11.90 s) was negligible. As a result, the total speedup and parallel efficiency for larger n was better than

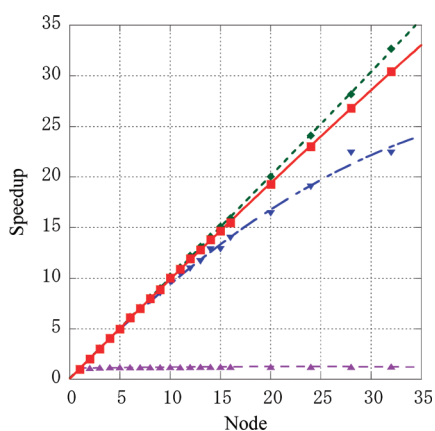


Figure 3. Speedup of the $[\text{PtCl}_6]^{2-}$ calculations. Inverted triangles (∇): 1-virtual part; diamonds (\blacklozenge): 2-virtual part; triangles (\blacktriangle): internal part; and squares (\blacksquare): total.

Table 5. p–s and p–p Excitation Energies of the Rn Atom (eV)

configuration	term	ΔJ	ref- CI	GMC- QDPT	ref (%) ^a	SAOP	ΔSCF (LDA)	expt
$6p^5(^2P_{3/2}^o)7s$	$^2[3/2]^o$	2	7.08	6.73	82.9	6.54	6.56	6.77
		1	7.26	6.93	83.4	6.74	6.72	6.94
$6p^5(^2P_{3/2}^o)7p$	$^2[1/2]$	1	8.30	8.22	83.4			8.21
		0	8.71	8.79	85.8			8.64
$6p^5(^2P_{3/2}^o)7p$	$^2[5/2]$	2	8.37	8.30	83.6			8.26
		3	8.53	8.50	84.0			8.43
$6p^5(^2P_{3/2}^o)7p$	$^2[3/2]$	1	8.58	8.55	83.9			8.46
		2	8.63	8.63	84.4			8.52
$6p^5(^2P_{1/2}^o)7s$	$^2[1/2]^o$	0	11.24	10.51	77.1	10.03	10.74	10.66
		1	11.29	10.52	79.2	10.11	10.61	10.79
$6p^5(^2P_{1/2}^o)7p$	$^2[1/2]$	1	12.56	11.98	81.7			
		0	12.65	12.09	80.1			
$6p^5(^2P_{1/2}^o)7p$	$^2[3/2]$	1	12.75	12.21	80.1			
		2	12.75	12.24	82.1			

^a The reference weight of the ground state was 87.2%.

the previous cases. The parallel efficiency for $n = 32$ in the present case was 0.95, which was better than in the $[\text{PtCl}_4]^{2-}$ (0.82) and $[\text{PtBr}_4]^{2-}$ (0.89) cases.

3.4. 6p–7p and 6p–7s Excitation Energies of the Radon Atom. The last example is the excitation energies of the radon atom. The target states were the 14 excited states resulting from 6p–7s and 6p–7p single excitations as well as the ground state. The reference space was of multireference singles and doubles (MRSD) type constructed from six electrons and 14 spinors (corresponding to 6p, 7s, and 7p orbitals). The determinants that spanned the reference CI space were generated from the DHF $(6p)^6(7s)^0(7p)^0$ configuration and selected according to their weights in the reference functions. All spinors were included in the perturbation calculation.

The computed results are summarized in Table 5 together with experimental values³³ and TDDFT [statistical average of the orbital model exchange–correlation potential (SAOP) and ΔSCF] results³⁴ for the 6p–7s excitations. For the lowest four 6p–7s excitations, experimental values of 6.77 eV ($\Delta J = 2$), 6.94 eV

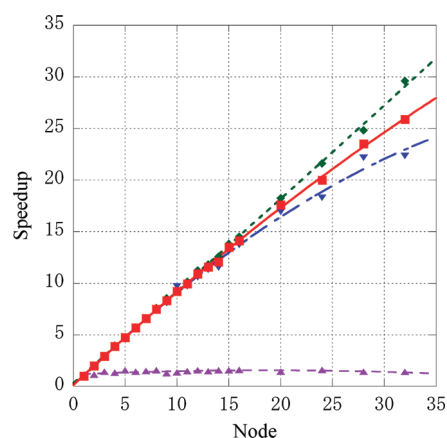


Figure 4. Speedup of the Rn atom calculations. Inverted triangles (∇): 1-virtual part; diamonds (\blacklozenge): 2-virtual part; triangles (\blacktriangle): internal part; and squares (\blacksquare): total.

($\Delta J = 1$), 10.66 eV ($\Delta J = 0$), and 10.79 eV ($\Delta J = 1$) are available. These excitation energies were computed to be 6.73, 6.93, 10.51, and 10.52 eV, respectively. The former two values were in very good agreement with the experimental values, whereas the latter two had larger errors. For the latter two states, the reference weights were relatively low, and the excitation energies at the reference space CI level were too large, which indicates that the reference functions were of low quality compared with those of the other excited states. This was because the main characteristics of these states were not well expressed using the MRSD-type reference space generated from the DHF configuration. The present results had better accuracy than the TDDFT results. The TDDFT with SAOP underestimated 6p–7s excitation energies by 0.68 eV at most, compared with the experimental values. In addition, the computed values for the six lowest 6p–7p excited states were also in good agreement with experimental data, and the maximum deviation was 0.15 eV.

Figure 4 is a plot of speedup ratios for the internal, 1-virtual, and 2-virtual terms. Speedup of the calculation for 2-virtual terms, which is the bottleneck of the calculation, was nearly proportional to the number of nodes. The parallel efficiency value for $n = 32$ was 0.81, which was somewhat worse than those for the $[\text{PtBr}_4]^{2-}$ and $[\text{PtCl}_6]^{2-}$ cases. The wall time for the 1-virtual terms (355.38 s) was close to that for the 2-virtual terms (469.09 s), and it was not proportional to the number of nodes, as in the previous cases. Consequently, the speedup and parallel efficiency values were a little worse than the previous cases.

4. CONCLUSION

We have developed a new, efficient parallel algorithm for four-component relativistic GMC-QDPT introducing Kramers symmetry. Our new algorithm has two advantages. The first advantage is speeding up the calculation, which is expected to be linear with the number of computational nodes n . The second is a reduction in the memory resources required to $1/n$ compared with single-node calculations. In addition, because of the Kramers restriction, the amount of memory required to store two-electron integrals, which are necessary for perturbation calculations, was reduced to three-eighths of the requirement for conventional GMC-QDPT calculations.

We applied our new scheme to calculation of the d–d excitation energies of the platinum halide complexes, $[\text{PtCl}_4]^{2-}$,

[PtBr₄]²⁻, and [PtCl₆]²⁻ and 6p–7s and 6p–7p excitation energies of the Rn atom. The present parallel algorithm had high efficiency, approximately proportional to the number of nodes. In GMC-QDPT calculations, the main bottleneck is calculation of the 2-virtual terms. Our new algorithm works linearly for these terms. Overall, the present algorithm showed high performance. In the case of platinum halide complexes, the calculated results were in good agreement with experimental values. We analyzed the wave function to evaluate our selection of reference functions. The reference weights were large enough for each state, and their deviations were small. Thus, these calculations have a good quality and balance. For the Rn atom, computational values of the two lower states of the 4 calculated 6p–7s excited states showed good agreement with the experimental values, while the computational values of 2 higher states had larger errors, and the reference weights were slightly smaller than for the lower two states.

AUTHOR INFORMATION

Corresponding Author

*E-mail: nakano@ccl.scc.kyushu-u.ac.jp.


REFERENCES

- (1) Johnson, W. R.; Idrees, M.; Sapirstein *J. Phys. Rev. A* **1987**, *35*, 3218–3226.
- (2) Ishikawa, Y. *Phys. Rev. A* **1990**, *42*, 1142–1150.
- (3) Quiney, H. M.; Grant, I. P.; Wilson, S. *J. Phys. B* **1990**, *23*, L271.
- (4) Blundell, S. A.; Johnson, W. R.; Sapirstein *J. Phys. Rev. Lett.* **1991**, *65*, 1411–1414.
- (5) Dyall, K. G. *Chem. Phys. Lett.* **1994**, *224*, 186–194.
- (6) Visscher, L.; Saue, T.; Nieuwpoort, W. C.; Fægri, K., Jr.; Groppen, O. *J. Chem. Phys.* **1993**, *99*, 6704–6715.
- (7) Watanabe, Y.; Matsuoka, O. *J. Chem. Phys.* **2002**, *116*, 9585–9590.
- (8) Eliav, E.; Kaldor, U. *J. Phys. Rev. A* **1994**, *49*, 1724–1729.
- (9) Visscher, L.; Dyall, K. G.; Lee, T. J. *Int. J. Quantum Chem., Symp.* **1995**, *29*, 411–419.
- (10) Fleig, T.; Jensen, H. J.; Olsen, J.; Visscher, L. *J. Chem. Phys.* **2006**, *124*, 104106.
- (11) Fleig, T.; Olsen, J.; Marian, C. M. *J. Chem. Phys.* **2001**, *114*, 4775–4790.
- (12) Fleig, T.; Olsen, J.; Visscher, L. *J. Chem. Phys.* **2003**, *119*, 2963–2971.
- (13) Fleig, T.; Sorensen, L. K.; Olsen, J. *Theor. Chem. Acc.* **2007**, *118*, 347–356.
- (14) Nakano, H.; Uchiyama, R.; Hirao, K. *J. Comput. Chem.* **2002**, *23*, 1166–1175.
- (15) Miyajima, M.; Watanabe, Y.; Nakano, H. *J. Chem. Phys.* **2006**, *124*, 044101.
- (16) Ebisuzaki, R.; Watanabe, Y.; Nakano, H. *Chem. Phys. Lett.* **2007**, *442*, 164–169.
- (17) Umeda, H.; Koseki, S.; Nagashima, U.; Schmidt, M. W. *J. Comput. Chem.* **2001**, *22*, 1243–1251.
- (18) Yanai, T.; Harrison, R. J.; Nakajima, T.; Ishikawa, Y.; Hirao, K. *Int. J. Quantum Chem.* **2007**, *107*, 1382–1389.
- (19) Yanai, T.; Nakajima, T.; Ishikawa, Y.; Hirao, K. *J. Chem. Phys.* **2001**, *114*, 6526–6538; **2002**, *116*, 10122–10128.
- (20) Abe, M.; Iikura, H.; Kamiya, M.; Nakajima, T.; Yanagisawa, S.; Yanai, T. *RELAD*; University of Tokyo: Tokyo, 2004.
- (21) Yanai, T.; Kamiya, M.; Kawashima, Y. et al. *UTCHEM*; University of Tokyo: Tokyo, 2004.
- (22) Koga, T.; Tatewaki, H.; Matsuoka, O. *J. Chem. Phys.* **2001**, *115*, 3561–3565. Koga, T.; Tatewaki, H.; Matsuoka, O. *J. Chem. Phys.* **2002**, *117*, 7813–7814.
- (23) Dyall, K. G. *Theor. Chem. Acc.* **2002**, *108*, 335–340. Dyall, K. G. *Theor. Chem. Acc.* **2006**, *115*, 441–447.
- (24) Sterzel, M.; Autschbach, J. *Inorg. Chem.* **2006**, *45*, 3316–3324.
- (25) Kroening, R. F.; Rush, R. M.; Martin, D. S., Jr.; Clardy, J. C. *Inorg. Chem.* **1974**, *13*, 1366–1373.
- (26) Granovsky, A. *Firefly*, version 7.1.G; Firefly Software: Austin, TX, 2009; <http://classic.chem.msu.su/gran/firefly/index.html>.
- (27) Patterson, H. H.; Godfrey, J. J.; Khan, S. M. *Inorg. Chem.* **1972**, *11*, 2872–2878.
- (28) Wang, F.; Ziegler, T. *J. Chem. Phys.* **2005**, *123*, 194102.
- (29) Kroening, R.; Rush, R. M.; Martin, D. S., Jr.; Clardy, J. C. *Inorg. Chem.* **1974**, *13*, 1366–1373.
- (30) Yoo, R. K.; Keiderling, T. A. *J. Phys. Chem.* **1990**, *94*, 8048–8055.
- (31) Jørgensen, C. K. *Acta Chem. Scand.* **1956**, *10*, 518–534.
- (32) Wang, F.; Ziegler, T. *J. Chem. Phys.* **2005**, *123*, 154102.
- (33) Moore, C. E. *Atomic Energy Levels, National Bureau of Standards Circular 467*; U.S. Government Printing Office: Washington, D.C., 1949, 1952, 1958.
- (34) Gao, J.; Liu, W.; Song, B.; Liu, C. *J. Chem. Phys.* **2004**, *121*, 6658–6666.

Concerted or Stepwise Mechanism? CASPT2 and LC-TDDFT Study of the Excited-State Double Proton Transfer in the 7-Azaindole Dimer

Xue-fang Yu, Shohei Yamazaki, and Tetsuya Taketsugu*

Division of Chemistry, Graduate School of Science, Hokkaido University, Sapporo 060-0810, Japan

 Supporting Information

ABSTRACT: Excited-state double proton transfer (ESDPT) in the 7-azaindole dimer is investigated using the complete active space second-order perturbation theory (CASPT2) method and the long-range corrected time-dependent density functional theory (LC-TDDFT) method. These methods are employed for geometry optimizations as well as single-point energy calculations of the excited-state potential energy profiles along the reaction paths. It is shown that three main reaction routes involving double proton transfer exist. In the first route, the ESDPT reaction takes place in the locally excited state through a single transition state following the concerted mechanism in which each proton-transfer process occurs simultaneously without forming any stable zwitterionic intermediate. The concerted ESDPT reaction is found to proceed asynchronously in C_s symmetry rather than synchronously in C_{2h} symmetry. In the second and third routes, on the other hand, the ESDPT reaction takes place following the stepwise mechanism in which each proton-transfer process occurs sequentially forming a neutral intermediate in the charge-transfer state. The calculated energy profiles of the three routes exhibit a lower barrier in the first route than in the other routes, suggesting that the ESDPT in the gas phase is likely to follow the asynchronous concerted mechanism at the lowest excitation energy.

1. INTRODUCTION

Proton-transfer reactions play essential roles in physics, chemistry, and biology.^{1–3} Among them, excited-state double proton transfer (ESDPT) in the 7-azaindole (7AI) dimer has been receiving particular attention (see ref 4 for a recent review), because this process can be taken as a model of the photoinduced mutation in DNA base pairs. Thus this process has been intensively studied experimentally^{4–38} and theoretically^{22,29,34,39–50} for more than 40 years. Through these studies, one major question has been put forward: Does the ESDPT follow a concerted mechanism or a stepwise mechanism? Figure 1 shows schematic pictures of the mechanisms of the ESDPT in the 7AI dimer from the normal dimer (ND) to the tautomer dimer (TD), derived from previous studies. In the concerted mechanism (Figure 1a), two protons are simultaneously transferred through a single transition state, without forming a stable intermediate for the single proton-transferred (SPT) component. The concerted DPT does not necessarily require synchronous motion of the two protons in which molecular structure of the dimer strictly keeps C_{2h} symmetry, but it can take asynchronous motion in which the two protons are transferred cooperatively in C_s symmetry, breaking the C_{2h} structure of the dimer.³⁶ In the stepwise mechanism (Figure 1b), on the other hand, a stable intermediate is formed by the first SPT from ND, followed by the second SPT to TD. This intermediate can have either a zwitterionic character or a neutral character.^{42,43,46,49} A zwitterionic intermediate is formed when the first SPT occurs in the locally excited (LE) state, while a neutral intermediate is formed when the first SPT occurs with transition to the charge-transfer (CT) state.

The mechanism of ESDPT in the 7AI dimer has been extensively discussed among several experimental groups, but it is still in controversy. Zewail and co-workers^{17,19,22,37} proposed the stepwise mechanism based on time-resolved spectra of the

7AI dimer in nonpolar and polar solvent as well as in the gas phase. The authors observed biexponential decay in the electronic spectra and assigned the faster and slower components of the decay to the first and second SPTs of the stepwise ESDPT, respectively. The decay time was found to strongly depend on solvent polarity, which was explained by the existence of a zwitterionic intermediate.³⁷ Castleman and co-workers^{20,23,24} supported the stepwise mechanism using the Coulomb explosion technique. On the other hand, Takeuchi and Tahara^{18,21,26,36} presented evidence of the concerted mechanism by examining the time-resolved fluorescence decay and its dependency on the excitation wavelength for the 7AI dimer in nonpolar solvent. These authors also detected the biexponential decay of fluorescence in ND as Zewail and co-workers did. However, they attributed the faster and slower components to internal conversion from the S_2 to S_1 state and the concerted ESDPT in the S_1 state, respectively, because the faster component was not observed at the lowest excitation energy. Sekiya, Sakota, and co-workers^{4,28,31–33,35} supported the concerted mechanism in terms of frequency- and time-resolved electronic spectra of the isolated 7AI dimer and its deuterated compounds. In particular, they concluded that the biexponential decay of fluorescence is not a proper evidence of the stepwise mechanism by using picosecond time-resolved resonance-enhanced multiphoton ionization (REMPI) spectroscopy.

Meanwhile, a large number of theoretical studies^{22,29,34,39–50} on the ESDPT in the 7AI dimer have been reported using semiempirical methods, ab initio methods, and density functional theory (DFT), but they have also exhibited different mechanisms. Several authors^{22,42,43,46} proposed the stepwise mechanism

Received: January 8, 2011

Published: March 16, 2011

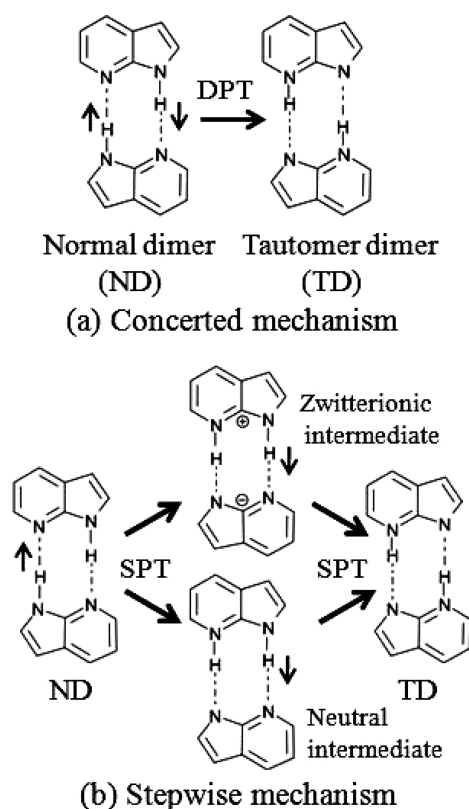


Figure 1. Two mechanisms of ESDPT in the 7AI dimer: (a) concerted mechanism and (b) stepwise mechanism.

via a zwitterionic intermediate or a neutral intermediate on the basis of the excited-state geometry optimizations with the configuration interaction singles (CIS) method. On the other hand, Catalán, Kasha, and co-workers^{29,34,44,45,47} concluded that the ESDPT follows the concerted mechanism in which the reaction path conserves C_{2h} symmetry by using the DFT and time-dependent DFT (TDDFT) methods with hybrid B3LYP functional. In 2006, Serrano-Andrés and Merchán⁴⁹ proposed the reaction paths of the ESDPT supporting the stepwise mechanism based on the single-point energy calculations with the complete active space second-order perturbation theory (CASPT2) method at the geometries optimized with the complete active space self-consistent-field (CASSCF) method. The authors reported that the stepwise reaction path through a zwitterionic intermediate should be more favorable than the concerted pathway in C_{2h} symmetry. However, the stepwise pathway with an intermediate minimum could not be reproduced in CASPT2 calculations by Nanbu and Sekiya (see ref 4). In addition, Ando and Kato⁵⁰ presented the reaction surface of the ESDPT supporting the concerted mechanism with very similar procedure as Serrano-Andrés and Merchán used: single-point energy calculations with the multireference second-order Møller–Plesset (MRMP2) method for the geometries optimized with the CASSCF method.

The present work aims to clarify the mechanism of the ESDPT reaction in the 7AI dimer thoroughly by means of more accurate ab initio and DFT calculations. For this purpose, the CASPT2 method^{51,52} and the long-range corrected TDDFT (LC-TDDFT) method^{53,54} are employed for geometry optimizations as well as single-point energy calculations along the excited-state reaction paths. Potential energy profiles are

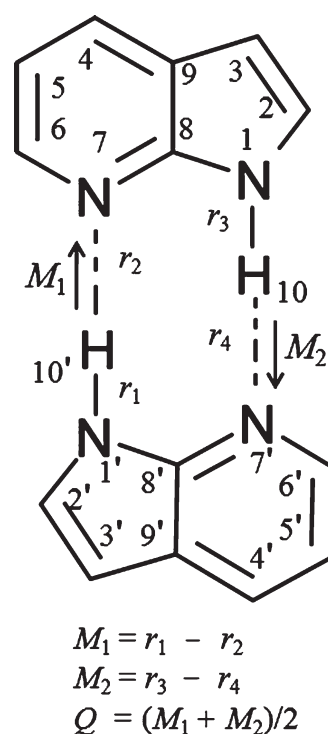


Figure 2. Definition of significant coordinates of ESDPT in the 7AI dimer. Labels for atoms are also given.

calculated for the ESDPT in the LE state as well as in the CT state in order to elucidate which mechanism in Figure 1 is the most favorable. Two major improvements in our approaches compared with other theoretical studies should be emphasized. First, the CASPT2 method is employed not only for single-point energy calculations but also for geometry optimizations in the excited states to take into account the dynamic electron correlation effect. Geometry optimizations with the CIS or CASSCF method may give less accurate structures due to the lack of dynamic correlation.⁵⁵ In the present work, analytical energy gradient⁵² is used for the CASPT2 geometry optimizations. Second, the long-range correction (LC) scheme is employed in TDDFT calculations. It is well-known that the TDDFT method with conventional exchange–correlation functionals, such as B3LYP, considerably underestimates the energies of CT states due to the lacking of nonlocal electron–electron exchange interaction.^{56,57} The LC scheme can overcome this disadvantage by modifying the long-range electron–electron interaction part of exchange functional by using the Hartree–Fock exchange integral. The reaction paths calculated at the CASPT2 and LC-TDDFT levels are expected to give a conclusive description of the ESDPT mechanism in the 7AI dimer.

2. COMPUTATIONAL DETAILS

Equilibrium geometries of ND and TD structures of the 7AI dimer in the ground state, referred to as ND_{S_0} and TD_{S_0} , respectively, and the geometry of the transition state between them, referred to as TS_{S_0} , were optimized by the Møller–Plesset second-order perturbation (MP2) method. For geometry optimization in the excited states, the CASPT2^{51,52} and LC-TDDFT^{53,54} methods were used. In the LC-TDDFT calculations, we employed the Becke 1988 exchange⁵⁸ and Lee–Yang–Parr correlation^{59,60} functional

with the long-range correction (LC-BLYP functional). Equilibrium geometries in the lowest singlet LE state of ND and TD structures, referred to as ND_{LE} and TD_{LE} , respectively, as well as the reaction path between ND_{LE} and TD_{LE} were determined by both the CASPT2 and LC-BLYP methods. The transition state between ND_{LE} and TD_{LE} in the LE state (referred to as TS_{LE}), minimum in the lowest singlet CT state (referred to as IN_{CT}), and the transition state between the LE and CT minima (referred to as TS_{CT}) were optimized by the LC-BLYP method. The geometry optimizations above were performed in C_s symmetry. Geometry optimizations with C_{2h} symmetry constraint were also performed by the LC-BLYP method for the minima of ND and TD structures in the first singlet excited state, referred to as $\text{ND}_{\text{S}_1}(C_{2h})$ and $\text{TD}_{\text{S}_1}(C_{2h})$, respectively, and the transition state between them, referred to as $\text{TS}_{\text{S}_1}(C_{2h})$. For the stationary points optimized with the MP2 method or the LC-BLYP method, normal-mode analysis was performed at the same computational level to check whether they are a minimum, a transition state, or a saddle point of higher order as well as to calculate the zero-point energy. Normal-mode analysis at the CASPT2 level could not be performed due to a huge computational cost.

Several reaction coordinates were introduced to describe a double proton transfer (DPT) process; see Figure 2. The coordinates $M_1 = r_1 - r_2$ and $M_2 = r_3 - r_4$, where r_1, r_2, r_3 , and r_4 are the length of $\text{N}1'-\text{H}10'$, $\text{N}7-\text{H}10'$, $\text{N}1-\text{H}10$, and $\text{N}7'-\text{H}10$ bond, respectively, characterize the transfer of each proton in $\text{N}-\text{H}\cdots\text{N}$ hydrogen bonds. The coordinate Q is defined by the average of M_1 and M_2 : $Q = (M_1 + M_2)/2$. The energy profiles along the reaction path from ND_{LE} to TD_{LE} were calculated as a function of the proton-transfer coordinate Q , in which the value of Q was fixed, and all other internal coordinates were optimized along the reaction path.

In geometry optimization at the CASPT2 level, the underlying state-averaged (SA) CASSCF wave function was generated with the active space of four electrons in four π orbitals, which are all localized on one of the monomers, and was averaged over the lowest three singlet states with equal weights (referred to as SA3-CASSCF(4,4)). Then single and double excitations from the reference space were taken into account by a perturbational treatment (referred to as CASPT2(4,4)) where 34 double-occupied orbitals were frozen to reduce the computational cost. A level shift with the parameter of 0.3 was employed for the CASPT2(4,4) calculations.⁶¹

For the CASPT2(4,4)-optimized geometries, single-point energies of the lowest 6 singlet states were calculated at the internally contracted CASPT2(12,12) level with 18 frozen core orbitals (1s orbitals of C and N atoms) and a level shift of 0.3, following SA6-CASSCF(12,12) calculations. The (12,12) active space includes 6 π orbitals of each monomer. The CASPT2-(12,12) method was also used for calculation of the vertical excitation energies at ND_{S_0} .

Along the reaction path in the LE state optimized with the LC-BLYP method, energies of the lowest 20 singlet excited states were calculated at the same level. To examine the effect of the LC scheme, the excited-state energies were also calculated by the conventional TDDFT method with the pure BLYP functional^{58–60} and hybrid B3LYP functional.^{62–64}

The Sapporo-DZP basis set^{65–68} was used for the calculations in the present work. One exception is for geometry optimization at the CASPT2(4,4) level along the reaction path, where polarization basis functions for hydrogen atoms other than the transferred ones were eliminated from Sapporo-DZP. CASPT2

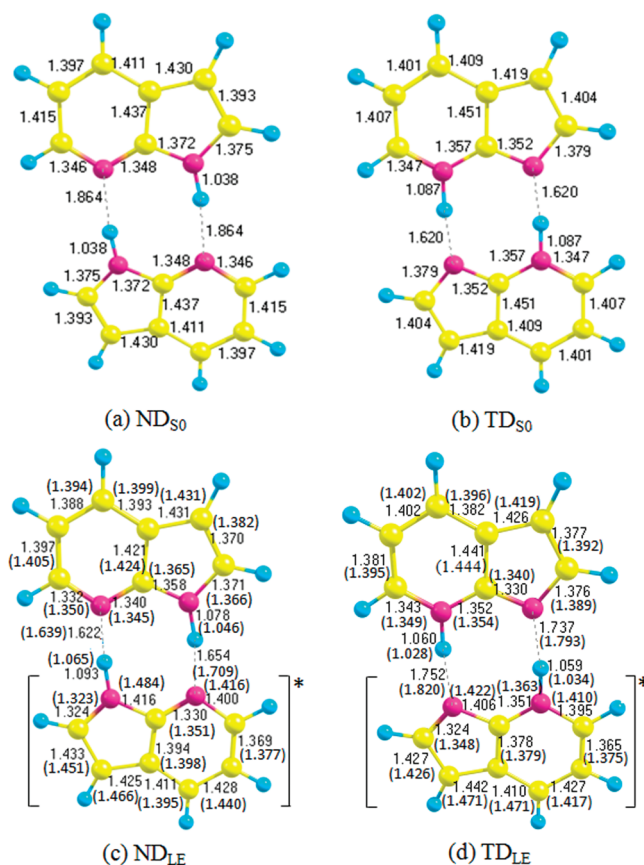


Figure 3. Equilibrium geometries of the 7AI dimer in the ground state, optimized at the MP2 level: (a) ND_{S_0} and (b) TD_{S_0} , and minimum-energy geometries of the 7AI dimer in the lowest LE state, optimized at the LC-BLYP level: (c) ND_{LE} and (d) TD_{LE} . In panels c and d, bond lengths obtained by CASPT2(4,4) optimizations are given in parentheses. Bond lengths are in angstroms. Monomer in []* indicates monomer-e.

calculations were performed with MOLPRO 2008.1,⁶⁹ while TDDFT calculations were performed with GAMESS.⁷⁰

3. RESULTS AND DISCUSSION

3.1. Equilibrium Geometries and Excitation Energies.

Figure 3 shows equilibrium structures in the ground and LE states, optimized by the MP2 method and the LC-TDDFT (LC-BLYP) method, respectively. For the LE minima, bond lengths obtained by the CASPT2(4,4) optimization are given in parentheses. The MP2-optimized geometries of ND_{S_0} and TD_{S_0} (Figure 3a and b, respectively) as well as TS_{S_0} (see the Supporting Information) belong to C_{2h} point group. The activation barrier from ND_{S_0} side is calculated as 15.0 kcal/mol, while the barrier from TD_{S_0} side is calculated as 0.9 kcal/mol. These barrier heights indicate that the reverse DPT in the ground state from TD_{S_0} to ND_{S_0} is very favorable and thus support the spectroscopic observation for TD in the ground state.⁷¹

As shown in Figure 3c and d, equilibrium structures in the lowest LE state (ND_{LE} and TD_{LE}) optimized at the LC-BLYP and CASPT2(4,4) levels are in C_s symmetry, that is, the two monomers of the 7AI dimer exhibit different structures. The excitation to the S_1 state at ND_{LE} and TD_{LE} is localized on one of the monomers. The excited monomer is referred to as monomer-e

(indicated by bracket and star in Figure 3c and d). The geometry of monomer-e is largely deviated by excitation, while the other monomer almost keeps the ground-state geometry (referred to as monomer-g). Thus the structure of the 7AI dimer deviates from C_{2h} symmetry at ND_{LE} and TD_{LE} . In the CASPT2(4,4) optimization, the four active orbitals for the underlying SA3-CASSCF(4,4) wave function are selected to be localized on monomer-e, see Figure 4. The active orbitals correspond to the highest occupied molecular orbital (HOMO), HOMO-1, the lowest unoccupied molecular orbital (LUMO), and LUMO+1 of monomer-e. In SA-CASSCF calculations, the ground and two LE states, the latter of which correspond to the 1L_a and 1L_b excitations⁷² localized on monomer-e, are averaged with equal weights. The 1L_a excitation is mainly composed by configuration

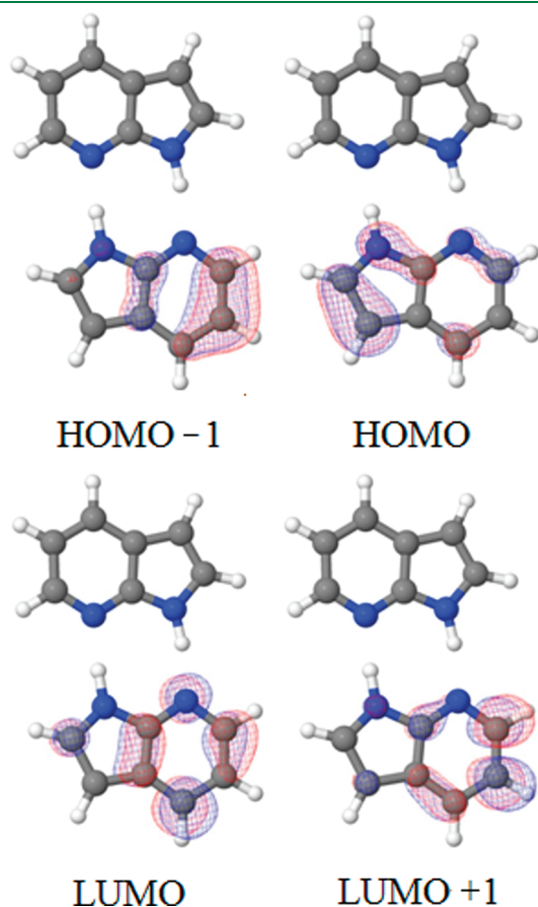


Figure 4. Active orbitals in CASPT2(4,4) calculations at ND_{LE} .

for the excitation from HOMO to LUMO, while the 1L_b excitation is mainly composed by configurations for the excitation from HOMO-1 to LUMO and the excitation from HOMO to LUMO+1. The LE states included in the SA3-CASSCF(4,4) calculations are referred to as $S_0-^1L_a$ and $S_0-^1L_b$, where the label $S_0-^1L_a$ ($S_0-^1L_b$) means that monomer-g is in the ground state, while monomer-e is in the 1L_a (1L_b) state. The lowest LE state is the $S_0-^1L_a$ state at ND_{LE} as well as at TD_{LE} optimized with the CASPT2(4,4) method. This is also true for ND_{LE} and TD_{LE} optimized with the LC-BLYP method.

Table 1 shows excitation and emission energies related to the ESDPT process in the 7AI dimer, including the vertical excitation energies for the lowest two singlet excited states at ND_{S_0} , the adiabatic excitation energies for the lowest LE state of ND, and the vertical emission energies for the lowest LE state at ND_{LE} and TD_{LE} , calculated at the LC-BLYP and the CASPT2(12,12) levels. For the vertical excitation of the 7AI dimer in C_{2h} symmetry, the lowest two excited states are 1^1B_u and 2^1A_g , and they are nearly degenerate at both the LC-BLYP and CASPT2 levels. The 1^1B_u and 2^1A_g states are optically allowed and forbidden, respectively, in C_{2h} symmetry. These states correspond to linear combinations of the configurations for the 1L_a local excitation on each monomer, i.e., $S_0-^1L_a$ and $^1L_a-S_0$. These findings are consistent with the results of recent theoretical studies.^{49,73} When the molecular structure is relaxed from C_{2h} to C_s symmetry, the excitation is localized in the $S_0-^1L_a$ and $^1L_a-S_0$ states, and the first excited state is thus stabilized.

The calculated excitation and emission energies in Table 1 qualitatively agree with experimental values. The vertical excitation energy of the 1^1B_u state is 4.74 eV at the LC-BLYP level and 3.98 eV at the CASPT2 level, corresponding to the wavelengths of 262 and 312 nm, respectively. These values are consistent with absorption spectra of ND in nonpolar solvent.^{19,36,38} The 2^1A_g state exhibits similar values of vertical excitation energy: 4.76 and 3.96 eV (261 and 313 nm) at the LC-BLYP and CASPT2 levels, respectively. Adiabatic excitation energy of ND is 4.30 eV (288 nm) at the LC-BLYP level and 3.93 eV (316 nm) at the CASPT2 level. The adiabatic excitation energy at the LC-BLYP level is corrected to be 4.16 eV (298 nm) when the zero-point energies of ND_{S_0} and ND_{LE} (calculated with the MP2 and LC-BLYP methods, respectively) are included. The calculated adiabatic excitation energies can be compared with the ionization spectra for jet-cooled ND (4.00 eV, 310 nm).^{12,15} Vertical emission energies for ND_{LE} and TD_{LE} are 3.93 and 2.97 eV (316 and 418 nm) at the CASPT2 level, while they are 3.63 and 2.02 eV (342 and 614 nm) at the LC-BLYP level. These results are also consistent with the fluorescence spectra of the 7AI dimer in nonpolar solvent with respect to the position of the peaks

Table 1. Vertical Excitation Energy, Adiabatic Excitation Energy, and Vertical Emission Energy Calculated at the LC-BLYP and CASPT2(12,12) Levels (eV)

method	vertical excitation		adiabatic excitation	vertical emission	
	ND_{S_0} (1^1B_u)	ND_{S_0} (2^1A_g)	$ND_{LE} \leftarrow ND_{S_0}$	ND_{LE}	TD_{LE}
LC-BLYP	4.74	4.76	4.30 (4.16) ^a	3.93	2.97
CASPT2(12,12)	3.98	3.96	3.93	3.63	2.02
expt	4.32, ^b 4.34 ^c		4.00 ^d	3.54 ^e	2.58, ^e 2.53 ^f

^aZero-point energy correction is implemented. ^bAbsorption band maximum in hexane.³⁶ ^cAbsorption band maximum in 3-methylpentane and ethylcyclohexane.³⁸ ^dBand origins for jet-cooled ND in ionization spectrum.^{12,15} ^eFluorescence band maximum in 3-methylpentane and ethylcyclohexane.³⁸ ^fFluorescence band maximum in hexane.³⁶

corresponding to the emission from ND and TD.^{5,19,36,38} The transition energies at the CASPT2 level tend to be lower than the experimental results, while the transition energies at the LC-BLYP level are a little higher than the experimental values. The excitation and emission energies given in Table 1 are also consistent with those calculated in previous theoretical studies.^{34,49}

Geometry optimization was also performed for ND and TD structures in the lowest excited state with constraint of C_{2h} symmetry [$ND_{S_1}(C_{2h})$ and $TD_{S_1}(C_{2h})$, respectively] at the LC-BLYP level. The resulting structures are shown in the Supporting Information. These stationary points were found to be higher in energy by 4.6 and 5.3 kcal/mol than ND_{LE} and TD_{LE} , respectively, in C_s symmetry. From the normal-mode analysis, both $ND_{S_1}(C_{2h})$ and $TD_{S_1}(C_{2h})$ are shown to have one imaginary frequency mode of b_u irreducible representation which breaks C_{2h} symmetry. This means that $ND_{S_1}(C_{2h})$ and $TD_{S_1}(C_{2h})$ correspond to the transition states in C_s symmetry connecting two ND_{LE} structures ($S_0-{}^1L_a$ and ${}^1L_a-S_0$) and two TD_{LE} structures ($S_0-{}^1L_a$ and ${}^1L_a-S_0$), respectively. The double-minimum feature of the potential energy surface is consistent with the weak coupling case of Frenkel-type exciton model discussed in ref 4.

From Figure 3, one can see that the bond lengths of the hydrogen-bonded $N-H\cdots N$ part are quite different between the S_0 minima and the LE minima. In particular, the $N\cdots H$ distances of ND (r_2 and r_4 , see Figure 2) are considerably shortened by excitation. With respect to ND_{S_0} optimized at the MP2 level and ND_{LE} optimized at the LC-BLYP level, the $N\cdots H$ distances r_2 and r_4 vary from 1.864 to 1.622 Å and 1.864 to 1.654 Å, respectively. These findings suggest that the strength of $N-H\cdots N$ hydrogen bonds of ND is enhanced in the excited state compared to the ground state. The DPT reaction from ND is therefore expected to become facilitated by the strengthened intermolecular hydrogen bonds in the LE state.

Compared to the CIS- and CASSCF-optimized geometries in previous studies,^{46,49} much shorter $N\cdots H$ bond lengths have been exhibited for the LE minimum geometries optimized by the LC-BLYP or CASPT2 method; taking the $N7\cdots H10'$ distance of ND_{LE} for example, the distance calculated at the LC-BLYP level (CASPT2 level) is shorter by 0.486 Å (0.469 Å) than the CASSCF-optimized value (2.108 Å).⁴⁹ It is also much shorter than CIS-optimized $N\cdots H$ distance of ca. 2.0 Å.^{42,46} This difference is caused by the limitation of the CIS and CASSCF methods which both do not take into account dynamic electron correlation. The S_0 minima optimized at the MP2 level also exhibit much shorter $N\cdots H$ distance than at the Hartree–Fock level and CASSCF level.^{42,49} The S_0 and LE minima optimized at the B3LYP level^{34,46} exhibit smaller difference of the $N\cdots H$ distance from those optimized at the MP2, LC-BLYP, and CASPT2 levels (difference is less than 0.14 Å).

For ND_{LE} , one can also notice that some bond lengths on the rings of monomer-e optimized with the LC-BLYP and CASPT2 methods are considerably different from those optimized with the CASSCF method.⁴⁹ In our LC-BLYP and CASPT2 calculations, the 1L_a (HOMO \rightarrow LUMO) local excitation of monomer-e induces significantly longer bond lengths of $N1'-C8'$ and $C2'-C3'$ as well as shorter bond lengths of $C8'-C9'$ and $C6'-C5'$ compared with the corresponding bonds of monomer-g (see Figure 3c). The longer bond lengths reflect the reduction of bonding character by excitation from HOMO, while the shorter bond lengths reflect the enhancement of bonding

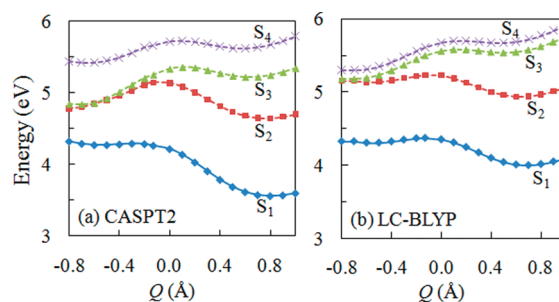


Figure 5. Potential energy profiles of the lowest four excited states of the 7AI dimer as functions of the reaction coordinate Q , calculated at the (a) CASPT2(12,12) and (b) LC-BLYP levels. Reaction paths in panels a and b are optimized for the lowest LE (S_1) state with the CASPT2(4,4) and LC-BLYP methods, respectively. Ground-state energy of ND_{S_0} is taken as zero. Full line with filled diamonds indicates energies of S_1 state whose geometries are optimized. Dashed lines show energies of higher LE states at the S_1 -optimized geometries.

character by excitation to LUMO (see Figure 4). The $C1'-C2'$ bond of monomer-e is also much shorter than $C1-C2$ bond of monomer-g, reflecting the reduction of antibonding character by excitation from HOMO. In the previous CASSCF calculation, on the other hand, $C8'-C9'$ and $C6'-C5'$ bond lengths of monomer-e are longer than the corresponding bonds of monomer-g, which may reflect the reduction of bonding character by excitation from HOMO-1 or excitation to LUMO+1. The difference suggests that the configuration of 1L_b local excitation has significant contribution to the S_1 state of ND_{LE} optimized at the CASSCF level. For TD_{LE} , our LC-BLYP and CASPT2 calculations as well as the previous CASSCF calculation predict the 1L_a local excitation of monomer-e, exhibiting longer bond lengths of $N1'-C8'$ and $C2'-C3'$ by excitation from HOMO and shorter bond lengths of $C8'-C9'$ and $C6'-C5'$ by excitation to LUMO (see Figure 3d).

3.2. Reaction Path of ESDPT in the Locally Excited State.

To understand the ESDPT mechanism of the 7AI dimer in the lowest LE state, the reaction path starting from ND_{LE} to TD_{LE} was determined at the CASPT2(4,4) level as well as at the LC-BLYP level along the reaction coordinate Q defined in Figure 2. Once the reaction path was determined, energies of the first and higher excited states were calculated along the CASPT2 and LC-BLYP reaction paths by the CASPT2(12,12) (for the lowest five excited states) and LC-BLYP (for the lowest 20 excited states) methods, respectively. Figure 5 shows the calculated energy profiles in the lowest four excited states where the energy of ND_{S_0} is taken as zero in each figure. At both the CASPT2 and LC-BLYP levels, the first excited state is dominated by the LE configuration labeled as $S_0-{}^1L_a$, where monomer-g is kept in the ground state and monomer-e is in the 1L_a excited state. At the LC-BLYP level, the second excited state corresponds to the ${}^1L_a-S_0$ state, where monomer-g is locally excited to the 1L_a state and monomer-e is in the ground state. The third and fourth excited states correspond to the local excitation to the 1L_b states of monomer-e and monomer-g, i.e., the $S_0-{}^1L_b$ and ${}^1L_b-S_0$ states, respectively. The fifth and higher excited states include ${}^1n-\pi^*$ states corresponding to the excitation from the lone-pair orbital of the N atoms as well as CT states where an electron is transferred between the two monomers. At the CASPT2 level, the second, third, and fourth excited states contain contributions of several electronic configurations including $S_0-{}^1L_a$, ${}^1L_a-S_0$,

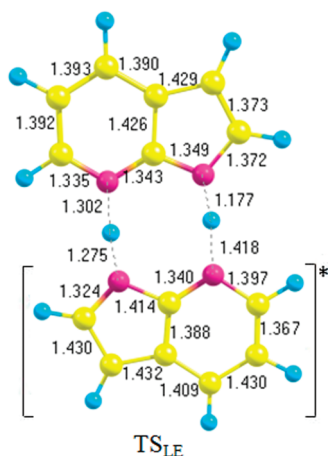


Figure 6. Geometry of transition state on DPT reaction path of the 7AI dimer in the lowest LE state (TS_{LE}), optimized at the LC-BLYP level. Bond lengths are in angstroms. Monomer in []* indicates monomer-e.

$S_0-{}^1L_b$, and ${}^1L_b-S_0$. The fifth excited state is described by a doubly excited configuration which can be labeled as ${}^1L_a-{}^1L_a$, where both monomers are simultaneously excited to the 1L_a state. It should be emphasized that the lowest four excited states along the reaction path are dominated by LE excitations at both CASPT2 and LC-BLYP levels and that no significant contribution of CT excitations is found. It is also worthy to mention that the LC-BLYP method reproduces the CASPT2 energy profiles in certain extent for the four LE states.

As shown in Figure 5, for the lowest LE state at both the CASPT2 and LC-BLYP levels, only one energy maximum has been found along the ESDPT reaction path connecting ND_{LE} and TD_{LE} , and no minimum for a stable zwitterionic intermediate could be located. The CASPT2 and LC-BLYP methods thus predict that the ESDPT in the LE state follows the concerted mechanism. Both methods show that activation barrier is very low for the concerted ESDPT from ND_{LE} to TD_{LE} . At the LC-BLYP level, the transition state TS_{LE} ($Q = -0.134$ Å) between ND_{LE} ($Q = -0.553$ Å) and TD_{LE} ($Q = 0.685$ Å) is located 1.5 kcal/mol (0.07 eV) higher in energy than ND_{LE} . When the zero-point correction is applied under the harmonic approximation, TS_{LE} is 1.9 kcal/mol (0.08 eV) lower in energy than ND_{LE} . This result suggests that the concerted ESDPT from ND_{LE} through TS_{LE} can proceed very efficiently. Figure 6 shows the optimized structure of TS_{LE} . The intrinsic reaction coordinate (IRC) calculations starting with TS_{LE} were also performed at the LC-BLYP level, which have confirmed that ND_{LE} and TD_{LE} are connected by the reaction path through TS_{LE} . Meanwhile, optimization of the transition-state geometry at the CASPT2 level could not be completed due to a huge computational cost. Nevertheless, the CASPT2 energy profile in Figure 5a shows a smaller energy barrier of 0.5 kcal/mol (0.02 eV) at $Q = -0.3$ Å between ND_{LE} ($Q = -0.5$ Å) and TD_{LE} ($Q = 0.8$ Å). The energy of TD_{LE} is calculated to be 16.4 kcal/mol (0.71 eV) and 7.0 kcal/mol (0.31 eV) lower than ND_{LE} at the CASPT2 and LC-BLYP levels, respectively, suggesting that the ESDPT process is exothermic.

The TS_{LE} geometry optimized at the LC-BLYP level in Figure 6 exhibits that the proton at the $N1'-H10'\cdots N7$ hydrogen bond is located almost in the middle of $N1'$ and $N7$ atoms ($r_1 = 1.275$ Å, $r_2 = 1.302$ Å, $M_1 = -0.027$ Å), while the

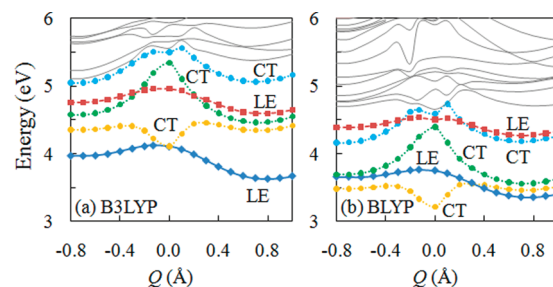


Figure 7. Potential energy profiles of the lowest five excited states of the 7AI dimer as functions of the reaction coordinate Q , calculated at the (a) B3LYP and (b) BLYP levels. Reaction paths are optimized for the lowest LE state with the LC-BLYP method. Ground-state energy of ND_{S0} is taken as zero. Full line with filled diamonds indicates the energies of the lowest LE state whose geometries are optimized. Dashed lines with filled squares indicate energies of other LE states, while dotted lines with filled circles indicate energies of CT states. Energies of sixth and higher excited states are also shown by thin solid lines.

proton at the $N1-H10\cdots N7'$ hydrogen bond is much closer to monomer-g than monomer-e ($r_3 = 1.177$ Å, $r_4 = 1.418$ Å, $M_2 = -0.241$ Å). A similar feature is found at the geometry of the energy maximum point of the CASPT2 potential energy curve ($Q = -0.3$ Å), where $r_1 = 1.189$ Å, $r_2 = 1.377$ Å, and $M_1 = -0.188$ Å in the $N1'-H10'\cdots N7$ hydrogen bond and $r_3 = 1.104$ Å, $r_4 = 1.516$ Å, and $M_2 = -0.421$ Å in the $N1-H10\cdots N7'$ hydrogen bond. These findings suggest an asynchronous DPT where the proton transfer from monomer-e to monomer-g along the former hydrogen bond is likely to be completed earlier than the proton transfer from monomer-g to monomer-e along the latter hydrogen bond. One point should be noted about the difference between asynchronous reaction (concerted reaction) and stepwise reaction. In asynchronous (concerted) reaction, two protons are transferred in different pace, but no stable SPT intermediate is formed. In stepwise reaction, on the other hand, a stable intermediate is formed after the first SPT. The asynchronous concerted mechanism has also been exhibited in ab initio molecular dynamics simulations of the excited-state hydrogen transfer in clusters of the 7AI monomer with one or two water molecules.⁷⁴ A transition state for synchronous ESDPT process with C_{2h} symmetry [$TS_{S1}(C_{2h})$] was also located by the LC-BLYP method (see the Supporting Information for the optimized structure). The energy of $TS_{S1}(C_{2h})$ is evaluated as 6.3 kcal/mol higher than ND_{LE} . The energy barrier indicates that the synchronous ESDPT process is less competitive to take place than the asynchronous ESDPT process, which exhibits a barrier of 1.5 kcal/mol. In addition, $TS_{S1}(C_{2h})$ exhibits two imaginary frequency modes, so this structure is characterized as a second-order saddle point. One of the imaginary frequency modes is an a_g mode corresponding to the symmetric proton-transfer motion, while the other is a b_u mode corresponding to in-plane motion of the rings, which breaks the C_{2h} symmetry. This finding means that the reaction path of the synchronous ESDPT from $ND_{S1}(C_{2h})$ to $TD_{S1}(C_{2h})$ through $TS_{S1}(C_{2h})$ is unstable with respect to the b_u imaginary frequency mode (see also Section 3.1).

As for ab initio methods, geometry optimization by the CASPT2 method in the present work has exhibited the concerted mechanism in the LE state and has not located any minima corresponding to a zwitterionic intermediate. Some recent ab initio studies using the CIS or CASSCF method for the geometry optimization, on the other hand, located an excited-state

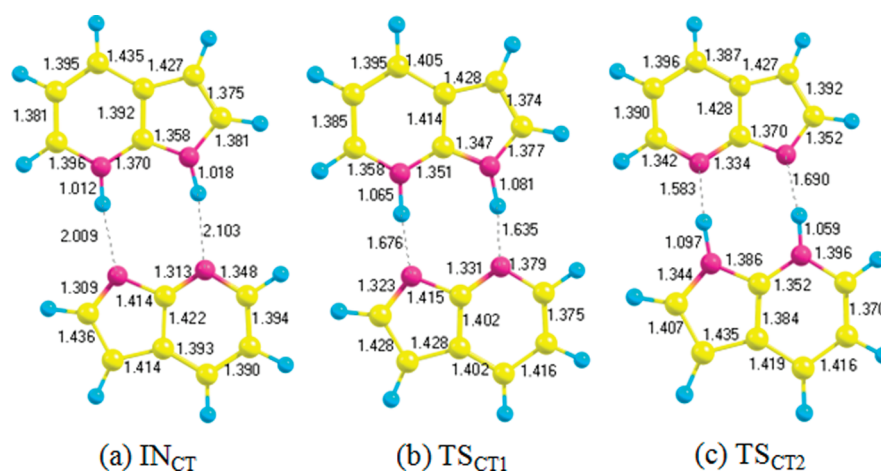


Figure 8. Geometries of (a) minimum in the lowest CT state (IN_{CT}) and (b and c) transition states on reaction paths of stepwise ESDPT through IN_{CT} (TS_{CT1} and TS_{CT2}) of the 7AI dimer, optimized at the LC-BLYP level. Bond lengths are in angstroms.

minimum for the zwitterionic intermediate.^{42,46,49} This discrepancy can be attributed to the lack of dynamic electron correlation in the latter methods. With respect to the CASSCF optimization, it should also be noticed that a minimum for the zwitterionic intermediate was located in one study,⁴⁹ while it could not be located in other studies.^{4,50} One possible explanation for this difference is that the intermediate minimum located in the former study may be artificial. In ref 49, the minimum of the zwitterionic intermediate is located in the S_2 state at the CASSCF level, while this electronic state changes to S_1 at the CASPT2 level. This change is caused by different contributions from dynamic electron correlation effects in the S_1 and S_2 states. In such a situation, shape of the CASPT2 potential energy surface may be significantly different from the CASSCF potential energy surface,⁷⁵ and it is therefore possible that the intermediate minimum at the CASSCF level disappears at the CASPT2 level.

We also applied the conventional TDDFT method with the B3LYP and BLYP functionals (without LC) to calculate the potential energy profiles of 20 lowest excited states along the reaction path optimized by the LC-BLYP method. Figure 7 shows the resulting potential energy curves of low-lying excited states. As is clearly shown here, the energies of CT states are drastically underestimated in both B3LYP and BLYP calculations compared with the LC-BLYP calculations. In the B3LYP calculations (Figure 7a), the lowest CT state corresponding to a charge transfer from monomer-e to monomer-g appears as the second excited state along the reaction path (yellow line). Moreover, the CT state is almost degenerate with the lowest LE state near the maximum point of the ESDPT potential energy curve ($Q = 0.0$ Å). In the BLYP calculations (Figure 7b), the first CT state is found to be below the lowest LE state around the ND_{LE} structure and exhibits considerably lower energy near the maximum point. The energy profiles at the B3LYP and BLYP levels are quite different from the LC-BLYP energy profiles calculated along the same reaction path (Figure 5b), where the first CT state is found to be much higher in energy than the lowest four LE states. At the B3LYP and BLYP levels, a large number of excited states are found within the range of less than 6 eV as shown in Figure 7. Note that in the LC-BLYP calculations only the four LE states are found in the same range (see Figure 5b).

The B3LYP method was also used for the optimization of the ESDPT reaction path. The resulting potential energy curve

(shown in the Supporting Information) exhibits a considerably deep minimum of the CT state near $Q = 0$ due to the large underestimation of the energies of this state. A similar behavior of the ESDPT potential energy curve at the B3LYP level was reported by Catalán and de Paz.³⁴ The authors calculated the potential energy profiles of the first excited state in C_s symmetry, resulting in significant energy lowering of intermediate structure and the dissociation of the 7AI dimer. Thus conventional TDDFT methods predict quite a different mechanism of the ESDPT in the 7AI dimer from the LC-TDDFT method as well as from the CASPT2 method.

3.3. Reaction Path in the Charge-Transfer State. The stepwise mechanism of the ESDPT via a neutral intermediate in the CT state was also investigated using the LC-BLYP method. Figure 8 shows the optimized geometries of stationary points relevant to the stepwise ESDPT. The minimum in the CT state for the neutral intermediate, labeled as IN_{CT} , has been found in a SPT structure ($M_1 = 0.997, M_2 = -1.085$ Å; see Figure 8a) and with the energy of 4.9 kcal/mol lower than ND_{LE} . At IN_{CT} , the first excited state is a CT state in which electron is transferred from HOMO of the monomer of proton donor to LUMO of the monomer of proton acceptor (see Figure 4 for HOMO and LUMO of the monomer), compensating the positive charge of the transferred proton. The lengths of $N1' \cdots H10'$ and $N7' \cdots H10$ hydrogen bonds (2.009 and 2.103 Å, respectively) are found to be much shorter than the calculated values in previous theoretical studies,^{42,46,49} due to the difference of computational methods used for the geometry optimization.

Transition states of SPT structure which may be accessed during the stepwise ESDPT through IN_{CT} have also been located. Figure 8b and c shows two transition-state structures optimized at the LC-BLYP level, referred to as TS_{CT1} ($M_1 = 0.611, M_2 = -0.544$ Å) and TS_{CT2} ($M_1 = -0.486, M_2 = 0.631$ Å), respectively. The transition state TS_{CT1} (TS_{CT2}) exhibits the energy which is 6.0 kcal/mol (6.9 kcal/mol) higher than ND_{LE} and 10.9 kcal/mol (11.8 kcal/mol) higher than IN_{CT} . When the zero-point energy correction is implemented, the energy difference between ND_{LE} and TS_{CT1} (TS_{CT2}) is 3.5 kcal/mol (4.3 kcal/mol) and the energy difference between IN_{CT} and TS_{CT1} (TS_{CT2}) is 7.3 kcal/mol (8.1 kcal/mol). The electronic structure of the S_1 state at these transition states is characterized as mixing of CT and LE configurations, although the monomer exhibiting

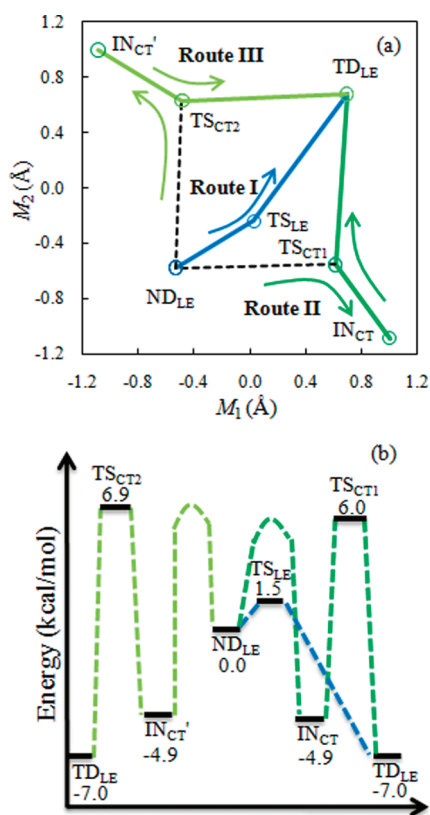


Figure 9. Schematic pictures of the reaction paths of the ESDPT in the 7AI dimer. (a) Reaction routes plotted in the (M_1, M_2) plane, where M_1 and M_2 are the reaction coordinates defined in Figure 2 and obtained by LC-BLYP optimization; and (b) energy diagram for reaction routes. Energies (relative to ND_{LE}) are calculated at the LC-BLYP level.

local excitation is different between TS_{CT1} and TS_{CT2} : The LE configuration at TS_{CT1} and TS_{CT2} corresponds to the 1L_a local excitation on the proton-donating monomer and the proton-accepting monomer, respectively. The character of the CT configuration at the transition states is the same as at IN_{CT} , where electron is transferred from HOMO of the proton-donating monomer to LUMO of the proton-accepting monomer. In other words, proton and electron are transferred from monomer-e to monomer-g at TS_{CT1} , while they are transferred from monomer-g to monomer-e at TS_{CT2} . The S_2 state at both TS_{CT1} and TS_{CT2} is found to be nearly degenerate with the S_1 state and also characterized as mixing of LE and CT configurations. This suggests that the TS_{CT} structures are located near the conical intersection and appear as a result of an avoided crossing of the S_1 and S_2 states.

The IRC calculations show that TS_{CT1} connects IN_{CT} and TD_{LE} while TS_{CT2} connects IN_{CT}' (inverted structure of IN_{CT} ; $M_1 = -1.085$, $M_2 = 0.997$ Å) and TD_{LE} . This result suggests that TS_{CT1} and TS_{CT2} can be accessed during the second SPT from IN_{CT} and IN_{CT}' to TD_{LE} . As to the first SPT process from ND_{LE} , IN_{CT} can be reached if proton and electron are transferred from monomer-e to monomer-g, while IN_{CT}' can be reached if the transfer occurs in the opposite direction. The second SPT from IN_{CT} and IN_{CT}' is characterized by the proton and electron transfer from monomer-g to monomer-e passing through TS_{CT1} and from monomer-e to monomer-g passing through TS_{CT2} , respectively. Unfortunately, the transition state connecting ND_{LE}

and IN_{CT} via an IRC path could not be found in the present calculations. We tried several transition-state searches starting from medium structures between ND_{LE} and IN_{CT} as well as between ND_{LE} and IN_{CT}' but all the optimizations led to TS_{CT1} or TS_{CT2} .

If the first SPT exhibits a barrier whose height is similar to the energy differences between TS_{CT1} and ND_{LE} (6.0 kcal/mol) and between TS_{CT2} and ND_{LE} (6.9 kcal/mol), then this process would be less likely to occur than the asynchronous concerted DPT in the LE state discussed in Section 3.2, which exhibits a barrier of 1.5 kcal/mol at the LC-BLYP level. Even if the first SPT exhibits a lower barrier, the stepwise ESDPT through IN_{CT} may still be less likely to be completed, because the second SPT from IN_{CT} (or IN_{CT}') to TD_{LE} along the IRC path through TS_{CT1} (or TS_{CT2}) exhibits a much higher barrier of 10.9 (or 11.8) kcal/mol. It may also be possible that the dimer which has arrived at IN_{CT} (or IN_{CT}') returns to the ground state through internal conversion resulting from small energy differences between the S_0 and S_1 states before the second SPT occurs.⁴⁹ This possibility has also been discussed for 2-aminopyridine dimer, another model of DNA base pairs,^{57,76} as well as DNA base pairs themselves (guanine-cytosine and adenine-thymine pairs).^{77,78}

3.4. ESDPT Mechanism. In the preceding sections, three routes have been presented for the ESDPT reaction in the 7AI dimer:

- Route I: $ND_{LE} \rightarrow TS_{LE} \rightarrow TD_{LE}$ (concerted mechanism).
- Route II: $ND_{LE} \rightarrow IN_{CT} \rightarrow TS_{CT1} \rightarrow TD_{LE}$ (stepwise mechanism).
- Route III: $ND_{LE} \rightarrow IN_{CT}' \rightarrow TS_{CT2} \rightarrow TD_{LE}$ (stepwise mechanism).

Figure 9 shows schematic pictures of the reaction routes. Figure 9a plots the value of the reaction coordinates M_1 and M_2 (defined in Figure 2) for minima and transition states relevant to each route, optimized with the LC-BLYP method. The respective points are connected with straight lines. The transition state connecting ND_{LE} to IN_{CT} or IN_{CT}' could not be found in the present calculations, so the pathways for the SPT reactions among these minima are temporarily expressed by dashed lines connecting ND_{LE} to TS_{CT1} and TS_{CT2} for visibility. Figure 9b shows a diagram for the energy profile of each route calculated at the LC-BLYP level.

In Route I, the 7AI dimer in the lowest LE state tautomerizes from ND_{LE} to TD_{LE} by concerted DPT, surmounting an energy barrier of 1.5 kcal/mol at TS_{LE} (the barrier height is estimated to be 0.5 kcal/mol at the CASPT2 level). The ESDPT reaction proceeds asynchronously, where the C_{2h} symmetry of the dimer is broken. The synchronous ESDPT conserving C_{2h} symmetry exhibits higher potential energies than the asynchronous ESDPT in C_s symmetry accompanying the unstable reaction path with respect to the in-plane mode which breaks the C_{2h} symmetry. In Routes II and III, on the other hand, two SPTs occur sequentially. For Route II, the first SPT from ND_{LE} to IN_{CT} forms a neutral intermediate by proton and electron transfer from monomer-e to monomer-g. The second SPT from IN_{CT} to TD_{LE} occurs through TS_{CT1} , overcoming an energy barrier of 10.9 kcal/mol by proton and electron transfer from monomer-g to monomer-e. TS_{CT1} is 6.0 kcal/mol higher in energy than ND_{LE} . Similarly, for Route III, the first SPT from ND_{LE} to IN_{CT}' forms a neutral intermediate. However, the direction of the proton and electron transfer is opposite to Route II, from monomer-g to monomer-e. For the second SPT from IN_{CT}' to TD_{LE} , the dimer has to

overcome an energy barrier of 11.8 kcal/mol at TS_{CT2} by proton and electron transfer from monomer-e to monomer-g. TS_{CT2} is 6.9 kcal/mol higher in energy than ND_{LE} .

According to the energy profiles of the three routes, it seems that the asynchronous concerted mechanism through Route I is the most likely to be followed. In the case of the gas phase, at least, one can say that the ESDPT in the 7AI dimer proceeds with the concerted mechanism at the lowest excitation energy. The stepwise mechanism through Routes II or III may also be possible at higher excitation energies.

In the concerted mechanism, the ESDPT in the 7AI dimer is a single-step process. This is consistent with the single-exponential decay of time-resolved electronic spectra of ND at the lowest excitation energy observed in the gas phase³³ as well as in a nonpolar solvent.³⁶ Presumably, the biexponential decay of other spectra in the two phases can be explained by combination with another process at higher excitation energy. For the gas phase, ESDPT after vibronic excitation of intermolecular stretching mode may correspond to the faster component of the biexponential decay, as Sakota et al.³³ proposed. In the case of nonpolar solvent, Takeuchi and Tahara³⁶ assigned the faster decay of fluorescence spectrum to internal conversion from higher to the first electronic excited state.

In the experiments on deuterated compounds of the 7AI dimer, moderate difference was found in the ESDPT rate constant of two isotopomers (one of them is deuterated on the NH group of monomer-e, while the other one is deuterated on the NH group of monomer-g).^{32,37} This kinetic isotope effect supports the asynchronous concerted mechanism in C_s symmetry, where the two SPTs proceed in different pace. If the ESDPT process were a synchronous process where C_{2h} symmetry is strictly retained, the two rate constants would be the same.

Photochemical behavior of the 7AI dimer can be considerably different in polar solvent. Based on time-resolved fluorescence spectra, Kwon and Zewail³⁷ proposed that the rate of ESDPT is significantly dependent on the polarity of solvent. On the other hand, Catalán⁷⁹ concluded that the 7AI molecule does not form a hydrogen-bonded dimer in polar solvent using steady-state absorption and emission spectroscopy. If the ESDPT occurs also in polar solvent with the asynchronous concerted mechanism presented in this work, then electrostatic interaction between solute and solvent molecules is expected to largely affect the reaction rate, because the dimer exhibits nonzero dipole moment along the reaction path owing to C_s symmetry. This is in clear contrast to the synchronous mechanism in C_{2h} symmetry, where the dipole moment of the dimer remains to be zero. To elucidate the mechanism of the excited-state process in polar solvent, however, additional theoretical studies would be necessary taking into account the effect of solvation on the excited-state potential energy surfaces, which is beyond the scope of the present work.

4. CONCLUSIONS

The present paper has shown that the ESDPT in the isolated 7AI dimer is likely to follow the concerted mechanism in terms of excited-state potential energy profiles calculated by the CASPT2 and LC-TDDFT methods. Three routes have been presented for the ESDPT reaction paths. In Route I, the DPT reaction takes place in the LE state through a single transition state following the concerted mechanism. The concerted ESDPT process has been found to occur asynchronously in C_s symmetry rather than synchronously in C_{2h} symmetry. The energy barrier for this

asynchronous ESDPT process is estimated to be 1.5 kcal/mol at the LC-TDDFT level and 0.5 kcal/mol at the CASPT2 level. In Routes II and III, on the other hand, the DPT reaction takes place via a neutral intermediate in the CT state following the stepwise mechanism, in which two SPTs occur sequentially. Routes II and III show opposite directions of the charge transfer between monomer-e and monomer-g. The second SPT from the intermediate requires to overcome a barrier of 10.9 and 11.8 kcal/mol in Routes II and III, respectively; these barriers are much higher than the barrier for Route I. Meanwhile, the zwitterionic intermediate presented in previous studies could not be reproduced in any routes.

One important point is that ab initio and DFT methods predict the same mechanism of the ESDPT in the LE state by improving each method. In the present calculations, the CASPT2 and LC-TDDFT methods predict very similar structures of stationary points as well as similar potential energy profiles in low-lying excited states for Route I. It is shown that the conventional TDDFT methods without the LC scheme drastically underestimate the energies of CT states, resulting in a wrong picture that many CT states are included in low-lying excited states along Route I. On the other hand, the LC-BLYP method exhibits no CT states in low-lying excited states along Route I and thus reproduces the CASPT2 results very well. Furthermore, CASSCF optimizations of ND_{LE} are found to predict qualitatively different structure from the CASPT2 and LC-TDDFT optimizations, presumably due to mixing of the $S_0-^1L_a$ and $S_0-^1L_b$ configurations. This discrepancy indicates that involvement of dynamic electron correlation is decisive in geometry optimizations for a reliable description of ESDPT process in the 7AI dimer.

This work sheds some new light on the long-lasting question about the mechanism of the ESDPT in the 7AI dimer. To explore the present study further, on-the-fly dynamics simulations of the DPT process in the gas phase and polar and nonpolar solvents would be an interesting topic of future studies.

■ ASSOCIATED CONTENT

S Supporting Information. Optimized structures of TS_{S0} , $ND_{S1}(C_{2h})$, $TD_{S1}(C_{2h})$, and $TS_{S1}(C_{2h})$; and potential energy curve of the S_1 state optimized at the B3LYP level. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: take@sci.hokudai.ac.jp.

■ ACKNOWLEDGMENT

This work was supported by a Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science, and Technology. The computations were performed using the Research Center for Computational Science, Okazaki, Japan. X.Y. thanks the China Scholarship Council, and S.Y. thanks the Japan Society for the Promotion of Science for Research Fellowships for Young Scientists. The authors wish to acknowledge contributions from Ai Saito and Daisuke Nakamura in the initial stage of the present study.

REFERENCES

- (1) *Proton Transfer in Hydrogen-Bonded Systems*; Bountis, T., Ed.; Plenum: New York, 1992.
- (2) Douhal, A.; Lahmani, F.; Zewail, A. H. *Chem. Phys.* **1996**, *207*, 477.
- (3) *Hydrogen-Transfer Reactions*; Hynes, J. T., Klinman, J. P., Limbach, H.-H., Schowen, R. L., Eds.; Wiley-VCH: Weinheim, Germany, 2007.
- (4) Sekiya, H.; Sakota, K. *J. Photochem. Photobiol. C* **2008**, *9*, 81.
- (5) Taylor, C. A.; El-Bayoumi, M. A.; Kasha, M. *Proc. Natl. Acad. Sci. U.S.A.* **1969**, *63*, 253.
- (6) Ingham, K. C.; Abu-Elgheit, M.; El-Bayoumi, M. A. *J. Am. Chem. Soc.* **1971**, *93*, 5023.
- (7) Ingham, K. C.; El-Bayoumi, M. A. *J. Am. Chem. Soc.* **1974**, *96*, 1674.
- (8) El-Bayoumi, M. A.; Avouris, P.; Ware, W. R. *J. Chem. Phys.* **1975**, *62*, 2499.
- (9) Hetherington, W. M., III; Micheels, R. H.; Eissenthal, K. E. *Chem. Phys. Lett.* **1979**, *66*, 230.
- (10) Waluk, J.; Bulska, H.; Pakuła, B.; Sepioł, J. *J. Lumin.* **1981**, *24/25*, 519.
- (11) Bulska, H.; Grabowska, A.; Pakuła, B.; Sepioł, J.; Waluk, J.; Wild, U. P. *J. Lumin.* **1984**, *29*, 65.
- (12) Fuke, K.; Yoshiuchi, H.; Kaya, K. *J. Phys. Chem.* **1984**, *88*, 5840.
- (13) Tokumura, K.; Watanabe, Y.; Itoh, M. *J. Phys. Chem.* **1986**, *90*, 2362.
- (14) Tokumura, K.; Watanabe, Y.; Udagawa, M.; Itoh, M. *J. Am. Chem. Soc.* **1987**, *109*, 1346.
- (15) Fuke, K.; Kaya, K. *J. Phys. Chem.* **1989**, *93*, 614.
- (16) Share, P.; Pereira, M.; Sarisky, M.; Repinc, S.; Hochstrasser, R. M. *J. Lumin.* **1991**, *48/49*, 204.
- (17) Douhal, A.; Kim, S. K.; Zewail, A. H. *Nature* **1995**, *378*, 260.
- (18) Takeuchi, S.; Tahara, T. *Chem. Phys. Lett.* **1997**, *277*, 340.
- (19) Chachivilis, M.; Fiebig, T.; Douhal, A.; Zewail, A. H. *J. Phys. Chem. A* **1998**, *102*, 669.
- (20) Folmer, D. E.; Poth, L.; Wisniewski, E. S.; Castleman, A. W., Jr. *Chem. Phys. Lett.* **1998**, *287*, 1.
- (21) Takeuchi, S.; Tahara, T. *J. Phys. Chem. A* **1998**, *102*, 7740.
- (22) Fiebig, T.; Chachivilis, M.; Manger, M.; Zewail, A. H.; Douhal, A.; Garcia-Ochoa, I.; de La Hoz Ayuso, A. *J. Phys. Chem. A* **1999**, *103*, 7419.
- (23) Folmer, D. E.; Wisniewski, E. S.; Hurley, S. M.; Castleman, A. W., Jr. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 12980.
- (24) Folmer, D. E.; Wisniewski, E. S.; Castleman, A. W., Jr. *Chem. Phys. Lett.* **2000**, *318*, 637.
- (25) Catalán, J.; Kasha, M. *J. Phys. Chem. A* **2000**, *104*, 10812.
- (26) Takeuchi, S.; Tahara, T. *Chem. Phys. Lett.* **2001**, *347*, 108.
- (27) Catalán, J. *J. Phys. Chem. A* **2002**, *106*, 6738.
- (28) Sakota, K.; Hara, A.; Sekiya, H. *Phys. Chem. Chem. Phys.* **2004**, *6*, 32.
- (29) Catalán, J.; Pérez, P.; del Valle, J. C.; de Paz, J. L. G.; Kasha, M. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 419.
- (30) Catalán, J. *Phys. Chem. Chem. Phys.* **2004**, *6*, 4467.
- (31) Sakota, K.; Sekiya, H. *J. Phys. Chem. A* **2005**, *109*, 2718.
- (32) Sakota, K.; Sekiya, H. *J. Phys. Chem. A* **2005**, *109*, 2722.
- (33) Sakota, K.; Okabe, C.; Nishi, N.; Sekiya, H. *J. Phys. Chem. A* **2005**, *109*, 5245.
- (34) Catalán, J.; de Paz, J. L. G. *J. Chem. Phys.* **2005**, *123*, 114302.
- (35) Sekiya, H.; Sakota, K. *Bull. Chem. Soc. Jpn.* **2006**, *79*, 373.
- (36) Takeuchi, S.; Tahara, T. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 5285.
- (37) Kwon, O.-H.; Zewail, A. H. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 8703.
- (38) Catalán, J. *J. Phys. Chem. A* **2010**, *114*, 5666.
- (39) Pechenaya, V. I.; Danilov, V. I. *Chem. Phys. Lett.* **1971**, *11*, 539.
- (40) Catalán, J.; Pérez, P. *J. Theor. Biol.* **1979**, *81*, 213.
- (41) Waluk, J.; Bulska, H.; Grabowska, A.; Mordziński, A. *Nouv. J. Chim.* **1986**, *10*, 413.
- (42) Douhal, A.; Guallar, V.; Moreno, M.; Lluch, J. M. *Chem. Phys. Lett.* **1996**, *256*, 370.
- (43) Guallar, V.; Batista, V. S.; Miller, W. H. *J. Chem. Phys.* **1999**, *110*, 9922.
- (44) Catalán, J.; del Valle, J. C.; Kasha, M. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 8338.
- (45) del Valle, J. C.; Kasha, M.; Catalán, J. *Int. J. Quantum Chem.* **2000**, *77*, 118.
- (46) Moreno, M.; Douhal, A.; Lluch, J. M.; Castaño, O.; Frutos, L. M. *J. Phys. Chem. A* **2001**, *105*, 3887.
- (47) Catalán, J.; Pérez, P.; del Valle, J. C.; de Paz, J. L. G.; Kasha, M. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 5799.
- (48) Gelabert, R.; Moreno, M.; Lluch, J. M. *J. Phys. Chem. A* **2006**, *110*, 1145.
- (49) Serrano-Andrés, L.; Merchán, M. *Chem. Phys. Lett.* **2006**, *418*, 569.
- (50) Ando, K.; Kato, S. Construction of The Potential Function for the Excited State Double Proton Transfer Reaction in 7-Azaindole Dimer. In 13th International Congress of Quantum Chemistry, Helsinki, Finland, June 22–27, 2009; p 197.
- (51) Celani, P.; Werner, H.-J. *J. Chem. Phys.* **2000**, *112*, 5546.
- (52) Celani, P.; Werner, H.-J. *J. Chem. Phys.* **2003**, *119*, 5044.
- (53) Tawada, Y.; Tsuneda, T.; Yanagisawa, S.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2004**, *120*, 8425.
- (54) Chiba, M.; Tsuneda, T.; Hirao, K. *J. Chem. Phys.* **2006**, *124*, 144106.
- (55) Page, C. S.; Olivucci, M. *J. Comput. Chem.* **2003**, *24*, 298.
- (56) Dreuw, A.; Weisman, J. L.; Head-Gordon, M. *J. Chem. Phys.* **2003**, *119*, 2943.
- (57) Sobolewski, A. L.; Domcke, W. *Chem. Phys.* **2003**, *294*, 73.
- (58) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.
- (59) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
- (60) Miehlisch, B.; Savin, A.; Stoll, H.; Preuss, H. *Chem. Phys. Lett.* **1989**, *157*, 200.
- (61) Roos, B. O.; Andersson, K. *Chem. Phys. Lett.* **1995**, *245*, 215.
- (62) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
- (63) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623.
- (64) Hertwig, R. H.; Koch, W. *Chem. Phys. Lett.* **1997**, *268*, 345.
- (65) Noro, T.; et al. *Segmented Gaussian Basis Set*; Quantum Chemistry Group: Sapporo, Japan; <http://setani.sci.hokudai.ac.jp/sapporo/>. Accessed February 24, 2011.
- (66) Yamamoto, H.; Matsuoka, O. *Bull. Univ. Electro-Commun.* **1992**, *5*, 23.
- (67) Noro, T.; Sekiya, M.; Koga, T. *Theor. Chem. Acc.* **1997**, *98*, 25.
- (68) Noro, T.; Sekiya, M.; Koga, T. *Theor. Chem. Acc.* **2003**, *109*, 85.
- (69) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M.; et al. *MOLPRO*, version 2008.1; University College Cardiff Consultants Limited: Wales, U.K., 2008.
- (70) 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., Jr. *J. Comput. Chem.* **1993**, *14*, 1347.
- (71) Ishikawa, H.; Yabuguchi, H.; Yamada, Y.; Fujihara, A.; Fuke, K. *J. Phys. Chem. A* **2010**, *114*, 3199.
- (72) Platt, J. R. *J. Chem. Phys.* **1949**, *17*, 484.
- (73) Serrano-Andrés, L.; Merchán, M.; Borin, A. C.; Stålring, J. *Int. J. Quantum Chem.* **2001**, *84*, 181.
- (74) Kina, D.; Nakayama, A.; Noro, T.; Taketsugu, T.; Gordon, M. S. *J. Phys. Chem. A* **2008**, *112*, 9675.
- (75) Serrano-Andrés, L.; Merchán, M. *J. Mol. Struct. (THEOCHEM)* **2005**, *729*, 99.
- (76) Schultz, T.; Samoylova, E.; Radloff, W.; Hertel, I. V.; Sobolewski, A. L.; Domcke, W. *Science* **2004**, *306*, 1765.
- (77) Sobolewski, A. L.; Domcke, W.; Hättig, C. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 17903.
- (78) Perun, S.; Sobolewski, A. L.; Domcke, W. *J. Phys. Chem. A* **2006**, *110*, 9031.
- (79) Catalán, J. *Nature Precedings* **2009**, <http://hdl.handle.net/10101/npre.2009.3089.1>; [hdl:10101/npre.2009.3089.1](http://hdl.handle.net/10101/npre.2009.3089.1).

New Interaction Parameters for Oxygen Compounds in the GROMOS Force Field: Improved Pure-Liquid and Solvation Properties for Alcohols, Ethers, Aldehydes, Ketones, Carboxylic Acids, and Esters

Bruno A. C. Horta,^{†,*} Patrick F. J. Fuchs,^{‡,§,||} Wilfred F. van Gunsteren,[†] and Philippe H. Hünenberger^{†,*}

[†]Laboratory of Physical Chemistry, ETH Zürich, CH-8093 Zürich, Switzerland

[‡]INSERM UMR-S665, DSIMB, Paris, France

[§]Université Paris-Diderot, UFR Sciences du Vivant, Paris, France

^{||}Institut National de Transfusion Sanguine, Paris, France

ABSTRACT: A new parameter set (S3A6_{OXY}) is developed for the GROMOS force field, that combines reoptimized parameters for the oxygen-containing chemical functions (alcohols, ethers, aldehydes, ketones, carboxylic acids, and esters) with the current biomolecular force field version (S3A6) for all other functions. In the context of oxygen-containing functions, the S3A6_{OXY} parameter set is obtained by optimization of simulated pure-liquid properties, namely the density ρ_{liq} and enthalpy of vaporization ΔH_{vap} , as well as solvation properties, namely the free energies of solvation in water ΔG_{wat} and in cyclohexane ΔG_{che} , against experimental data for 10 selected organic compounds, and further tested for 25 other compounds. The simultaneous refinement of atomic charges and Lennard-Jones interaction parameters against the four mentioned types of properties provides a single parameter set for the simulation of both liquid and biomolecular systems. Small changes in the covalent parameters controlling the geometry of the oxygen-containing chemical functions are also undertaken. The new S3A6_{OXY} force-field parameters reproduce the mentioned experimental data within root-mean-square deviations of 22.4 kg m⁻³ (ρ_{liq}), 3.1 kJ mol⁻¹ (ΔH_{vap}), 3.0 kJ mol⁻¹ (ΔG_{wat}), and 1.7 kJ mol⁻¹ (ΔG_{che}) for the 35 compounds considered.

I. INTRODUCTION

Molecular dynamics (MD) simulation represents a powerful tool for investigating the properties of molecular systems relevant in physics, chemistry, and biology.^{1–4} The usefulness of this method in the context of condensed-phase systems results in particular from a favorable trade-off between model resolution and computational cost. Although classical atomistic models represent an approximation to quantum mechanics (QM), they can still provide a realistic description of molecular systems at spatial and temporal resolutions on the order of 0.1 nm and 1 fs, respectively, while their computational cost remains tractable at present for system sizes and time scales on the order of 10 nm and 100 ns, respectively. These scales are sufficient to enable in many cases (i) an appropriate description of solvation, by explicit treatment of the solvent molecules within a sufficiently large solvation range; (ii) a reliable calculation of thermodynamic properties *via* statistical mechanics; and (iii) a direct comparison with experimental data, namely structural, thermodynamic, transport, and dynamic observables measured on similar spatial and temporal scales.

In classical MD simulations, the atomic coordinates and velocities are propagated in time by integrating Newton's equations of motion, possibly thermostatted and barostatted,⁵ and reformulated in a discretized form.^{6–8} To this purpose, the forces acting on each atom at a given time step are calculated on the basis of an empirical potential-energy function, also called a force field. A force field is a parametric function of the coordinates of all atoms, and its specification requires the choice of a functional form and of the associated parameter set. The

functional form is typically defined by a sum of bonded and nonbonded terms designed to model certain types of physical interactions. The bonded terms are intramolecular and generally depend each on a single internal coordinate defined by a limited set of covalently bonded atoms, e.g., bond stretching, bond-angle bending, improper-dihedral-angle distortion, and dihedral-angle torsion. The nonbonded terms account for through-space interactions and generally depend each on a single interatomic distance, e.g., pairwise electrostatic and van der Waals interactions. The parameter set consists of a list of constants involved in the evaluation of the different force-field terms, e.g., reference internal-coordinate values, force constants, torsional-potential multiplicities and phase shifts, atomic charges, and pairwise van der Waals interaction parameters.

From a broad perspective, one may distinguish between two main classes of force fields, involving different scopes and design strategies. On one hand, molecular mechanics or spectroscopic force fields, e.g., CFF,^{9,10} CVFF,¹¹ MM3,^{12–14} and MM4,¹⁵ mainly aim at an accurate description of molecular properties in the gas phase, e.g., geometries, energies, and vibrational properties. They usually involve few atom types, a complex functional form, e.g., including anharmonicities and couplings in the bonded terms, largely automated parametrization procedures, and parameters mainly derived on the basis of spectroscopic measurements and QM calculations in the gas phase. These force fields typically focus on intramolecular

Received: November 8, 2010

Published: March 25, 2011

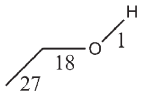
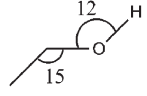
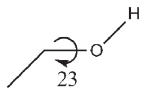
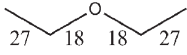
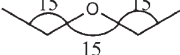
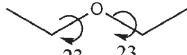
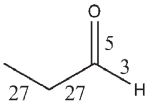
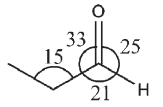
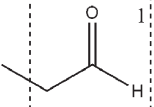
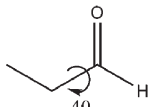
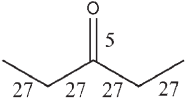
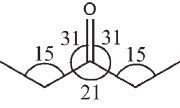
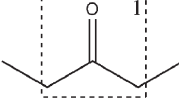
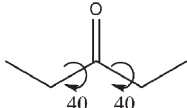
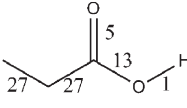
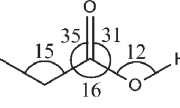
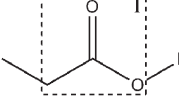
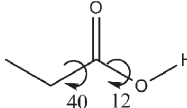
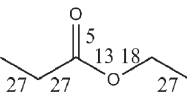
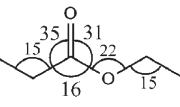
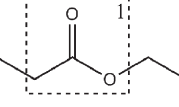
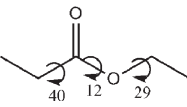
	bond type	bond-angle type	improper-dihedral-angle type	dihedral-angle type
alcohols			—	
ethers			—	
aldehydes				
ketones				
carboxylic acids				
esters				

Figure 1. Bonded interaction types (bond stretching, bond-angle bending, improper-dihedral-angle distortion, and dihedral-angle torsion) used in the 53A6_{OXY} parameter set. The numbering refers to the type codes of the 53A5 and 53A6 parameter sets.³⁴ The corresponding parameters are also listed in Table 3.

rather than on intermolecular interactions. Due to the often simplistic representation of long-range nonbonded interactions and solvation effects, their application is usually restricted to the conformational analysis of small molecules, as a classical alternative to more expensive QM calculations. On the other hand, condensed-phase or biomolecular force fields, e.g., CHARMM,^{16–19} AMBER,^{20–23} OPLS,^{24–27} and GROMOS,^{7,28–38} mainly aim at the description of (bio)molecules in solution. They usually involve a larger number of atom types, accounting for atoms in different chemical environments, a simpler functional form, empirical parametrization procedures, a more extensive use of parameter-combination and transferability assumptions, and, in particular for OPLS and GROMOS, parameters derived mainly from experimental spectroscopic and thermodynamic data concerning liquids and solutions. These force fields focus on the description of torsional-angle properties, nonbonded interactions, and solvation effects. For this reason, they are able to capture the main physics underlying the properties of condensed-phase systems such as solids, liquids, solutions, and solvated (bio)molecules.

The GROMOS force field belongs to the second category. It has been widely used in simulations of liquids,^{39,40} crystals,^{41,42} and solutions,^{43,44} and in the context of biomolecular systems, e.g., peptides^{45,46} and proteins,^{47,48} nucleic acids,^{35,49} carbohydrates,^{50,51} and lipids.^{52,53}

The main principles underlying the construction of this force field can be summarized as follows: (i) united atom representation of aliphatic CH, CH₂, CH₃, and CH₄ groups;^{28,30,31} (ii) quartic bond-stretching (since 1996), cosine-harmonic bond-angle bending (since 1996), cosine-series dihedral-angle torsion, and harmonic improper-dihedral-angle distortion terms; (iii) Lennard-Jones representation of the van der Waals interactions; (iv) mean-field representation of electronic polarization effects *via* enhanced atomic charges appropriate for condensed-phase (polar) environments (see also refs 54–57 for recent developments concerning explicit polarization); (v) nonbonded exclusion of first and second covalent neighbors; (vi) van der Waals interaction reduction for third covalent neighbors using a special set of parameters; (vii) van der Waals interaction adjustment distinguishing between non-hydrogen-bonding, uncharged hydrogen-bonding, and charged hydrogen-bonding interactions; (viii) application of a geometric-mean combination rule⁷ for van der Waals interaction parameters of particular sets of atoms; and (ix) freely adjustable atomic partial charges; i.e., these charges are not determined by the atom type.

The main principles underlying the parametrization strategy of the force field can be summarized as follows: (i) steady but controlled parameter refinement over the years based on small molecule data and avoiding unnecessary complexity increases;

Table 1. Summary of the Main GROMOS Force-Field Versions since the Original 37C4 Parameter Set

GROMOS force field nomenclature	year of release	main improvements	references
37C4	1987	GROMOS87 force field	ref 60
43A1	1996	GROMOS96 force field	refs 7, 28
43A2	2000	reparametrization of the dihedral-angle potentials for <i>n</i> -alkanes	ref 30
45A3	2001	improved description of liquid aliphatic hydrocarbons and alkane–water systems	ref 31
45A4	2005	improved description of lipids, nucleic acids, and carbohydrates together with slight readjustment in the choice of vdW interaction types	refs 32, 35, 36, 87
53A5	2004	reoptimization of the polar functional groups (vdW and charges) for pure liquids	ref 34
53A6	2004	reoptimization of the polar functional groups (charges only; vdW identical to 53A5) based on hydration free energies	ref 34
53A6 _{OL3/Chiu}	2009	improved parameters for phosphatidylcholine bilayers	ref 38
53A6 _{OXY}	2011	reoptimization of nonbonded interaction parameters for oxygen compounds (+ slight readjustments of reference bond lengths and angles within oxygen functions)	present work

(ii) principal focus on torsional-angle properties, nonbonded interactions, and solvation effects, most relevant for the description of condensed-phase and biomolecular systems; (iii) calibration involving in the first place primary experimental data, i.e., observed thermodynamic and spectroscopic properties concerning small molecules in the condensed phase; (iv) an assumption of transferability, justifying the application of parameters calibrated for small molecules to corresponding fragments within larger molecules; (v) compatibility with the simple-point-charge (SPC) water model;⁵⁸ and (vi) compatibility with reaction-field electrostatics⁵⁹ based on an effective long-range cutoff distance of 1.4 nm (since 1996).

The original 37C4 version of the GROMOS force field⁶⁰ (see refs 7, 34, 61 for a historical overview) has been progressively refined and extended, still maintaining the compatibility with the original functional form, except for the change^{7,8,29} to quartic bond stretching, cosine-harmonic bond-angle bending, and reaction field electrostatics in 1996, and the SPC water model.⁵⁸ The main consecutive versions of the force field are listed in Table 1. These were validated by simulations of liquids and solvated biomolecules.^{44,62,63} A complete description of the functional form of the GROMOS force field can be found in refs 7, 8, 29, and 34. The most recent published force-field parameter sets are labeled 53A5 and 53A6 and have been described by Oostenbrink et al.³⁴

The latest force-field version includes two distinct parameter sets, exclusively differing in the atomic partial charges within polar functional groups. Set 53A5, which was parametrized to reproduce thermodynamic properties, e.g., densities and enthalpies of vaporization, of pure liquids, is recommended for the simulation of organic liquids and liquid mixtures. Set 53A6, in which the atomic partial charges were readjusted so as to reproduce hydration free energies of polar amino acid analogs, is recommended for the simulations of biomolecular systems. The decision of providing two distinct parameter sets³⁴ resulted from the apparent impossibility of reproducing pure-liquid properties and hydration free energies simultaneously with a sufficient accuracy. This incompatibility may be an unavoidable consequence of the mean-field treatment of electronic polarization, rendering the derivation of parameters appropriate for both highly polar (aqueous solution) and less polar (pure organic liquids) environments impossible. However, two considerations may soften this statement. First, the derivation of set 53A6 only involved a readjustment of the charges, not of the van der Waals interaction parameters. Thus, this set may be viewed as a compromise combining charges appropriate for high-polarity

environments with van der Waals interaction parameters appropriate for lower-polarity environments. For this reason, it has been recommended, rather than set 53A5, for the simulation of systems where the partitioning of polar functional groups between polar and nonpolar environments is of relevance, e.g., biomolecules. This compromise is, however suboptimal, compared to allowing for the flexibility that would be offered by a simultaneous refinement of the two types of nonbonded interaction parameters. Second, the derivation of the common van der Waals interaction parameters of the two sets only involved the adjustment of the repulsive (C_{12}) coefficients of the Lennard-Jones interaction, not of the corresponding dispersive (C_6) coefficients. The reason for the latter choice was the qualitative connection existing between pairwise dispersive interactions and the electronic polarizabilities of the involved atoms, e.g., through the Slater–Kirkwood expression,⁶⁴ which suggests that the C_6 parameters should not be freely adjustable. However, considering that the original GROMOS C_6 parameters⁷ have been derived on the basis of gas-phase atomic polarizabilities, which might differ from the “effective” polarizabilities of atoms in molecular environments, limited adjustments of these parameters might be physically justified and used to further enhance the agreement with the target data.

The goal of the present work is to consistently reoptimize the nonbonded interaction parameters of set 53A6, in order to derive a new parameter set reconciling the reproduction of experimental data concerning pure organic liquids as well as aqueous and nonaqueous solvation properties within a reasonable accuracy, namely, as measured by root-mean-square deviations, on the order of a few percent (about 1–3%) in terms of densities and of $k_B T$ (2.5 kJ mol⁻¹ at room temperature) in terms of energetic properties. This is done by allowing more force-field parameters to be optimized than was the case in the calibration of the 53A5 and 53A6 versions of the force field. At present, solely oxygen compounds are considered, including the most common chemical functions of the elements C, H, and O, namely, alcohols, ethers, aldehydes, ketones, carboxylic acids, and esters. The refinement only affects the atomic partial charges within the above functions and the parameters for van der Waals interactions involving oxygen atoms. A slight readjustment in the reference values (but not force constants) of the bond-stretching and bond-angle-bending terms within these functions is also undertaken, so as to improve the description of their geometries. The resulting parameter set is referred to as 53A6_{OXY}.

Table 2. List of Oxygen Compounds Considered in This Study^a

chemical function	code	name	calibration ^b	ϵ_{rf}	$\kappa_{\text{T}} [10^{-4} (\text{kJ mol}^{-1} \text{nm}^{-3})^{-1}]$
alcohols	MTL	methanol		33.5	12.48
	ETL	ethanol	×	24.0	11.53
	PPL	propanol	×	20.0	10.26
	BTL	butanol		17.7	9.42
	PTL	pentanol		15.1	8.84
	HXL	hexanol		13.0	8.24
	HPL	heptanol		11.5	7.94 ^c
	OTL	octanol		10.1	7.64
	2PPL	propan-2-ol		19.1	13.32
	2BTL	butan-2-ol		16.7	9.42 ^c
	2PTL	pentan-2-ol		13.8	8.84 ^c
	3PTL	pentan-3-ol		13.4	8.84 ^c
	CHXL	cyclohexanol		16.4	8.24 ^c
	2M2P	2-methylpropan-2-ol		11.5	9.42 ^c
	2M2B	2-methylbutan-2-ol		5.7	8.84 ^c
ethers	DME	methoxymethane	×	6.2	8.00 ^c
	DEE	ethoxyethane	×	4.2	8.00 ^c
	MPH	1-methoxypropane		4.2 ^c	8.00 ^c
	DXE	1,2-dimethoxyethane		7.3	8.00 ^c
aldehydes	EAL	acetaldehyde	×	21.1	8.00 ^c
	PAL	propionaldehyde		18.4	8.00 ^c
	BAL	butyraldehyde		13.4	8.00 ^c
ketones	PPN	propanone	×	20.8	13.24
	BTN	butanone	×	17.7	11.88
	2PN	pentan-2-one		15.4	10.92
	3PN	pentan-3-one		16.6	10.92 ^c
	2HN	hexan-2-one		14.5	10.12
	3HN	hexan-3-one		14.5 ^c	10.12 ^c
acids	ACA	acetic acid	×	6.2	9.17
	PPA	propionic acid		3.4	9.29
	BTA	butyric acid		2.9	9.29 ^c
esters	EAE	ethylacetate	×	6.0	8.98 ^d
	PAE	propylacetate	×	5.6	8.67 ^d
	BAE	butylacetate		5.1	8.39 ^d
	MPE	methylpropionate		6.0	8.98 ^c

^a The acronyms adopted in this article and the official IUPAC names are provided. The values of the reaction-field static relative dielectric permittivity (ϵ_{rf}) used in the simulations are also listed and were taken from experimental results,⁸⁸ when available. The values of the isothermal compressibility κ_{T} used for the pressure scaling in the simulations are also provided and were taken from experimental results,⁶⁵ when available. ^b Compounds used in the calibration. ^c Not available experimentally and chosen on the basis of experimental data for similar compounds. ^d Not available experimentally and approximated by the corresponding adiabatic compressibility.

II. COMPUTATIONAL DETAILS

II.1. Parametrization Strategy. The 35 oxygen compounds considered in the present study are listed in Table 2, along with corresponding acronyms used in the article. With the exception of DME (boiling point⁶⁵ $T_{\text{b}} = 254$ K), all are in the liquid phase under standard conditions, i.e. at 1 bar and 298.15 K. Some of these compounds, indicated by the symbol “×”, were used in the parameter calibration, while others were only used to test the transferability of the reoptimized parameters. For example, ETL and PPL were selected to calibrate the alcohol parameters, while the 13 remaining alcohols were used for validation only.

The thermodynamic observables against which the parametrization was performed are (i) the density ρ_{liq} of the pure liquid, (ii) the enthalpy of vaporization ΔH_{vap} of the pure liquid, (iii) the solvation free energy ΔG_{wat} of the compound in water, and (iv) the solvation free energy ΔG_{che} of the compound in cyclohexane. The comparison between simulated and experimental values was performed under standard conditions, except for DME, where a temperature of 254 K was selected instead.

The improper-dihedral-angle distortion parameters, the dihedral-angle-torsion parameters, the definition of charge groups, and the special third-neighbor van der Waals interaction parameters were kept unaltered with respect to the S3A6 force field.³⁴

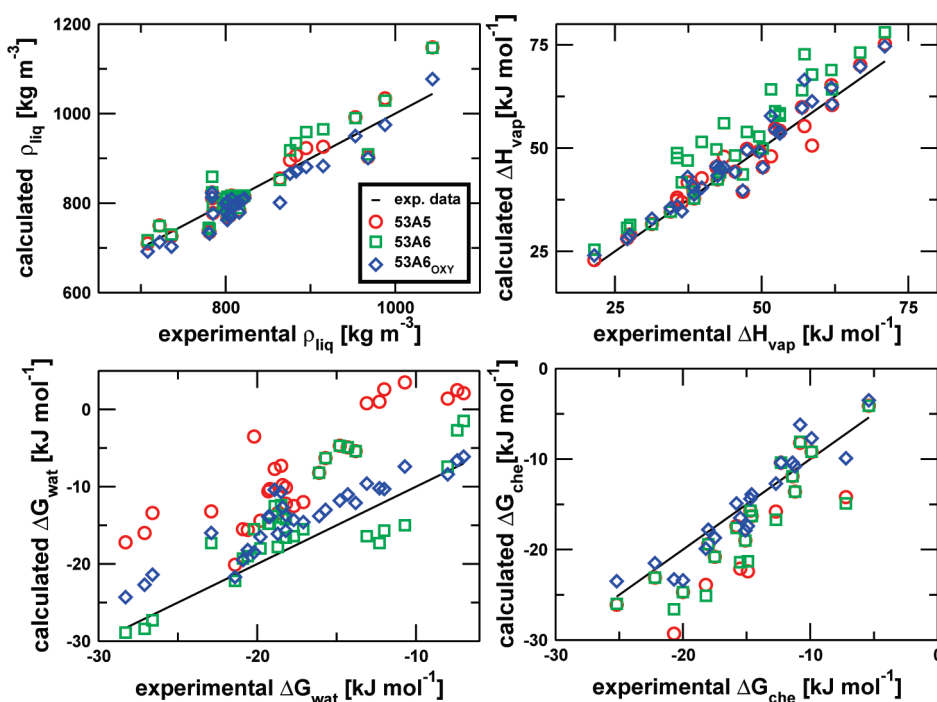


Figure 2. Comparison of simulated and experimental properties obtained with different GROMOS parameter sets (53A5, 53A6, and 53A6_{OXY}). The properties considered are the pure-liquid density ρ_{liq} and enthalpy of vaporization ΔH_{vap} as well as the solvation free energies in water ΔG_{wat} and in cyclohexane ΔG_{che} at 1 bar and 298.15 K. The straight line corresponds to a perfect agreement with experimental data.

On the other hand, the bond-stretching and bond-angle bending parameters were reassigned (from the list of existing covalent parameters), on the basis of the consideration of quantum-mechanically optimized molecular geometries (MP2/6-311++G**), as detailed in Figure 1 and Table 3. These calculations were performed using the Gaussian 03 program,⁶⁶ and the adjustment was performed on the basis of the optimized geometries of the compounds propanol, ethoxyethane, propionaldehyde, butanone, propionic acid, and ethylacetate.

The interaction parameters of the united aliphatic-carbon atoms (CH, CH₂, and CH₃) were not altered in the present work. These parameters have been previously optimized in the context of linear, branched, and cyclic alkanes against experimental pure-liquid properties, e.g., density and vaporization enthalpy, as well as hydration free energies.^{30,31} They were subsequently shown to reproduce very well other experimental properties, including solvation free energies in apolar^{44,67} and other polar solvents,⁴⁴ and conformational properties of hydrocarbons.³⁰ These parameters have been validated in the context of biomolecular simulations, and especially of lipids, e.g., in terms of melting temperatures for monoglycerides at different hydration levels.⁶⁸ Therefore, the quality of the united-aliphatic-atom parameters was not questioned, and these parameters were directly used as a starting point for the present reoptimization concerning oxygen-containing compounds.

In GROMOS, the van der Waals interactions are calculated according to a Lennard-Jones function, i.e., as

$$V^{\text{LJ}}(\mathbf{r}; C_{12}, C_6) = \sum_i \sum_{j>i} \left(\frac{C_{12,ij}}{r_{ij}^{12}} - \frac{C_{6,ij}}{r_{ij}^6} \right) \quad (1)$$

where r_{ij} is the (minimum-image) distance between two interaction sites i and j , and the parameters $C_{12,ij}$ and $C_{6,ij}$ are defined

following a geometric-mean combination rule

$$C_{12,ij} = C_{12,ii}^{1/2} C_{12,jj}^{1/2} \quad \text{and} \quad C_{6,ij} = C_{6,ii}^{1/2} C_{6,jj}^{1/2} \quad (2)$$

Up to three different $C_{12}^{1/2}$ parameters are actually defined for every type of atom, labeled here $C_{12,I}^{1/2}$, $C_{12,II}^{1/2}$, and $C_{12,III}^{1/2}$. The $C_{12,I}^{1/2}$, $C_{12,II}^{1/2}$, and $C_{12,III}^{1/2}$, normally corresponding to increasing repulsiveness, are used in the cases of non-hydrogen bonding, uncharged hydrogen-bonding, or charged hydrogen-bonding interactions, respectively. In the present work, all compounds considered are uncharged, and the $C_{12,III}^{1/2}$ is not relevant. The choice of the appropriate type of $C_{12}^{1/2}$ parameter for the interaction between two given types of atoms is defined by a combination matrix (refer to Table 8 of ref 34 for the 53A5/53A6 combination matrix, which is used in the present work).

The calibration procedure involved the refinement of the Lennard-Jones interaction parameters $C_6^{1/2}$, $C_{12,I}^{1/2}$, and $C_{12,II}^{1/2}$ associated with the atom types O (carbonyl oxygen), OA (alcohol or carboxylic acid oxygen), and OE (ether or ester oxygen), as well as of the partial charges of all of the atoms involved in a specific functional group. Distinct charge sets for alcohol, ether, aldehyde, ketone, carboxylic acid, and ester groups were used. The adjustment of the $C_6^{1/2}$ parameters was kept minimal, considering the qualitative connection that exists between this parameter and the electronic polarizability of the corresponding atom.⁶⁴

The calibration was performed by trial and error, steered to some extent by chemical intuition, and a roughly incremental approach. In particular, the Lennard-Jones interaction parameters refined for oxygen atoms in alcohols (OA), ethers (OE), and ketones (O) were used directly for aldehydes, carboxylic acids, and esters. The charge sets were refined separately for the six classes of compounds. During the parametrization of some of the alcohols and carboxylic acids, adequate parameters leading to

Table 3. Bonded Interaction Parameters (Bond Stretching, Bond-Angle Bending, Improper-Dihedral-Angle Distortion, and Dihedral-Angle Torsion) Used in the S3A6_{OXY} Parameter Set and Comparison with the Values Obtained from *ab Initio* Quantum-Mechanical Geometry Optimization

bond type code	K_b [10^6 kJ mol ⁻¹ nm ⁻⁴]	bond stretching		
		b_0 [nm]	b_{eq} (HF/6-31+G**) [nm]	b_{eq} (MP2/6-311++G**) [nm]
1	15.7	0.100	0.094 ^a /0.095 ^b	0.096 ^a /0.097 ^b
3	12.3	0.109	0.110 ^c	0.111 ^c
5	16.6	0.123	0.119 ^c /0.119 ^d /0.119 ^b /0.119 ^e	0.121 ^c /0.122 ^d /0.121 ^b /0.122 ^e
13	10.2	0.136	0.133 ^b /0.132 ^c	0.136 ^b /0.136 ^c
18	8.18	0.143	0.141 ^a /0.140 ^f /0.143 ^c	0.143 ^a /0.142 ^f /0.145 ^c
27	7.15	0.153	0.151–0.153 ^{a,b,c,d,e,f}	0.151–0.153 ^{a,b,c,d,e,f}
angle type code	K_θ [kJ mol ⁻¹]	θ_0 [deg]	bond-angle bending	
			θ_{eq} (HF/6-31+G**) [degree]	θ_{eq} (MP2/6-311++G**) [degree]
12	450	109.5	110.6 ^a /108.9 ^b	107.5 ^a /105.8 ^b
15	530	111.0	108.4–114.0 ^{a,b,c,d,e,f}	108.1–113.1 ^{a,b,c,d,e,f}
16	545	113.0	111.9 ^b /112.8 ^c	111.2 ^b /111.8 ^c
21	620	116.0	116.5 ^c /115.8 ^d	115.9 ^c /116.0 ^d
22	635	117.0	117.8 ^c	114.9 ^c
25	505	120.0	119.9 ^c	120.2 ^c
31	700	122.0	122.1 ^d /122.1 ^b /123.4 ^e	122.0 ^d /122.7 ^b /123.5 ^e
33	730	124.0	123.6 ^c	123.9 ^c
35	750	125.0	126.1 ^b /123.8 ^c	126.1 ^b /124.7 ^c
improper dihedral-angle type code	improper dihedral-angle distortion			
	K_ξ [kJ mol ⁻¹ deg ⁻²]	ξ_0 [deg]		
1	0.0510	0.0		
dihedral-angle torsion type code	dihedral-angle torsion			
	K_ϕ [kJ mol ⁻¹]	$\cos(\delta)$	m	
12	16.7	–1.0	2	
23	1.26	+1.0	3	
29	3.77	+1.0	3	
40	1.00	+1.0	6	

Molecules considered for the geometry optimization: ^a Propanol. ^b Propionic acid. ^c Propionaldehyde. ^d Butanone. ^e Ethylacetate. ^f Ethoxyethane.

the simultaneous reproduction of the experimental data for the four thermodynamic properties considered could not be found. In this case, a higher weight was given to ρ_{liq} , an intermediate weight to ΔH_{vap} and ΔG_{wat} and a lower weight to ΔG_{che} . The parametrization involved about 200 000 simulation runs for a total computing time of about 14 CPU years on a heterogeneous cluster mostly composed of quad-core AMD Opteron 8380 CPUs.

The final reoptimized nonbonded interaction parameters are provided in Tables 4 (charges) and 5 (Lennard-Jones), and compared with the corresponding values in the S3A5 and S3A6 force fields³⁴ as well as a variant, i.e., an ether set labeled S3A6_w.³⁷ The latter set involves modified Lennard-Jones interaction parameters for the oxygen atom type OE as well as a slightly different charge set for the ether function, and was refined specifically for ethers and polyethers against pure-liquid properties, hydration free energies, and conformational properties. The definition of the S3A6_{OXY} force field is completed by (i) the single-atom Lennard-Jones interaction parameters corresponding to the 50 unaltered atom types of the S3A6 force field, see Table VII in ref 34; (ii) the selection matrix for single-

atom parameter combinations defining non-hydrogen-bonding, uncharged hydrogen-bonding, and charged hydrogen-bonding pair Lennard-Jones parameters in the S3A6 force field, see Table VIII in ref 34; and (iii) the residue topology building blocks of the GROMOS force field,⁷ solely altered in the assignment of bond and bond-angle types (Figure 1 and Table 3) as well as atomic charges (Table 4) for the oxygen functions.

The details concerning the simulations required for the evaluation of ρ_{liq} , ΔH_{vap} , ΔG_{wat} and ΔG_{che} are provided in the following sections. Error estimates on the calculated quantities were evaluated by block averaging.¹ For the pure-liquid properties, ρ_{liq} and ΔH_{vap} , the errors were always below 1% and are not reported. For the solvation properties, ΔG_{wat} and ΔG_{che} , the total error was estimated by weighted summation of error contributions evaluated separately for the simulations at the successive λ -values (see section II.5).

II.2. Simulation Protocol. All simulations were performed using the GROMOS MD++ program,⁶⁹ release version 0.3.0. The GROMOS force-field parameter sets considered are S3A5 and S3A6,³⁴ the variant S3A6_w for ethers,³⁷ and the new

Table 4. Partial Charges of the Atoms within the Different Oxygen Functions in the GROMOS Force Field^a

chemical function	IAC	atom type	partial atomic charges			
			S3A5 [<i>e</i>]	S3A6 [<i>e</i>]	S3A6 _{variant} [<i>e</i>]	S3A6 _{OXY} [<i>e</i>]
alcohol	21	H	0.403	0.408		0.410
	3	OA	−0.611	−0.674		−0.700
	13–18	CHn	0.208	0.266		0.290
ether	4	OE	−0.324	−0.324	0.420 ^b	−0.580
	13–18	CHn	0.162	0.162	0.210 ^b	0.290
aldehyde	12	C				0.375
	20	HC				0.100
	1	O				0.475
ketone	1	O	−0.450	−0.450		−0.540
	12	C	0.450	0.450		0.540
acid	12	C	0.658	0.330		0.550
	1	O	−0.450	−0.450		−0.550
	3	OA	−0.611	−0.288		−0.410
	21	H	0.403	0.408		0.410
ester	13–18	CHn	0.160	0.266 ^c		0.290
	4	OE	−0.360	−0.069 ^c		−0.370
	1	O	−0.380	−0.450 ^c		−0.550
	12	C	0.580	0.253 ^c		0.630

^a The integer atom code (IAC) and the corresponding atom type are also indicated.³⁴ Each set of charges defines a single charge group. The S3A5 and S3A6 force fields³⁴ as well as the variant S3A6_W for ethers³⁷ and the parameter set S3A6_{OXY} derived in this work are shown. Empty entries correspond to parameters not available in the given set. ^b Variant labeled S3A6_W, introduced by Winger et al.³⁷, for the simulation of polyethers, to be used with OE parameters of the same set (Table 5). ^c Introduced by Chandrasekhar et al.³³ for the simulation of esters and possibly lipids (entry q1 in the second Table of ref 33).

Table 5. Lennard-Jones Interaction Parameters for the Three Oxygen Types Used in the Simulation of the Chemical Functions Considered in This Work^a

force field	IAC	atom type	C ₆ ^{1/2} [(kJ mol ^{−1} nm ⁶) ^{1/2}]	C ₁₂ ^{1/2} [10 ^{−3} (kJ mol ^{−1} nm ¹²) ^{1/2}]		
				I	II	III
S3A5(6)	1	O	0.04756	1.000	1.130	
				
	3	OA	0.04756	1.100	1.227	
	4	OE	0.04756	1.100	1.227	
S3A6 _W	4	OE	0.06313	2.148	1.227	1.748
				
S3A6 _{OXY}	1	O	0.04136	0.995	1.100	
				
	3	OA	0.04500	1.150	1.350	
	4	OE	0.04123	1.529	1.529	

^a The integer atom type code (IAC) and corresponding atom type³⁴ are indicated along with the single-atom Lennard-Jones interaction parameters C₆^{1/2} and C₁₂^{1/2} entering the GROMOS geometric-mean combination rule.⁷ The parameters C_{12,I}^{1/2}, C_{12,II}^{1/2}, and C_{12,III}^{1/2} refer to non-hydrogen-bonding, uncharged hydrogen-bonding, and charged hydrogen-bonding interactions. The indicated parameters correspond to the S3A5 and S3A6 force fields,³⁴ to a variant S3A6_W for ethers,³⁷ and to the parameter set S3A6_{OXY} derived in this work. Empty entries correspond to parameters not available in the given set.

parameter set S3A6_{OXY} developed in the present work. Note that in the case of ester compounds, the charge set corresponding to the S3A6 force field is the one developed by Chandrasekhar et al.

(entry q1 of the second table of ref 33). The solvent models considered are the SPC water model⁵⁸ and the GROMOS S3A6 cyclohexane model originally developed by Schuler et al.³¹

All simulations, including the gas-phase simulations, were carried out under periodic boundary conditions based on cubic computational boxes. Newton's equations of motion were integrated using the leapfrog scheme⁷⁰ with a time step of 2 fs. All bond lengths were constrained by application of the SHAKE procedure⁷¹ with a relative geometric tolerance of 10^{-4} . The temperature was maintained close to its reference value of 298.15 K (254 K for DME) by weak-coupling to external baths⁵ using a relaxation time of 0.1 ps. In all simulations, including the gas-phase simulations, distinct temperature baths were used for the translational and for the internal and rotational degrees of freedom of the molecules. Note that a joint coupling to a single temperature bath was used instead in ref 34 for both types of simulations, which explains the presence of small differences between the data reported therein and the present data for the 53A5 and 53A6 force fields (only significant for the gas-phase simulations). Except for the gas-phase simulations, which were performed at a constant volume, all simulations were carried out at constant pressure. The pressure was maintained close to its reference value of 1 bar by isotropic weak-coupling of the atomic coordinates and box dimensions to a pressure bath⁵ using a relaxation time of 0.5 ps. The isothermal compressibilities involved in the pressure coupling were set to 4.575×10^{-4} ($\text{kJ mol}^{-1} \text{nm}^{-3}$)⁻¹ for the simulations in water and to 11.2×10^{-4} ($\text{kJ mol}^{-1} \text{nm}^{-3}$)⁻¹ for the simulations in cyclohexane, or set equal to the experimental compressibility for the pure liquids (Table 2). The center of mass motion was removed every 100 ps. The nonbonded interactions were computed using a twin-range scheme,^{2,7} with short- and long-range cutoff distances set to 0.8 and 1.4 nm, respectively, and an update frequency of five time steps for the short-range pairlist and intermediate-range interactions. A reaction-field correction^{59,72} was applied to account for the mean effect of electrostatic interactions beyond the long-range cutoff distance, using relative dielectric permittivities of 61 for the simulations in water⁷³ and 6 for the simulations in cyclohexane, or set equal to the experimental permittivity value for the pure liquids (Table 2). The reaction-field self-term and excluded-atom-term contributions⁷⁴ to the energy, forces, and virial were included as described in ref 69.

II.3. Pure-Liquid Simulations. The initial coordinates for the simulations of the pure liquids were generated by randomly placing 512 molecules in cubic computational boxes of dimensions appropriate for the experimental density of the liquid. After energy minimization, the systems were equilibrated, first at constant volume (0.25 ns) and then at constant pressure (1 ns). The production runs at constant pressure used for the calculation of ρ_{liq} and ΔH_{vap} covered an additional 2 ns.

II.4. Gas-Phase Simulations. The estimation of the gas-phase intramolecular energies is required for the calculation of ΔH_{vap} . These gas-phase simulations were carried out as described in previous work,^{30,31,34} by simulating systems in which the individual molecules are initially placed very far apart from each other within a periodic box. The initial coordinates for these simulations were generated by randomly placing 512 molecules in cubic computational boxes of edge length ~ 750 nm, respecting a minimal intermolecular distance of 50 nm. This value is sufficient to ensure that the distances between atoms within different molecules do not exceed the long-range cutoff distance of 1.4 nm throughout the simulations. This setup mimics a gas-phase situation while providing more statistics compared to a single-molecule simulation, and permitting a coupling to a nonstochastic thermostat. The systems were equilibrated at

constant volume for 0.1 ns, followed by 0.2 ns of production. The enthalpy of vaporization ΔH_{vap} was calculated as the difference between the average potential energy per molecule in the gas-phase and pure-liquid simulations, expressed on a per mole basis and increased by RT , where R is the ideal-gas constant and T is the absolute temperature. This corresponds to a standard-state definition involving a reference pressure of 1 bar for both the gas-phase and liquid standard states, as recommended by IUPAC.⁷⁵ The term $M_{\text{liq}}\rho_{\text{liq}}^{-1}$, where M_{liq} is the molar mass of the compound, accounting for the volume-pressure enthalpy contribution of the pure liquid, is in all cases very small and was neglected.

II.5. Solvation Free-Energy Calculations. Simulations in water and cyclohexane were carried out in order to evaluate the corresponding solvation free energies ΔG_{wat} and ΔG_{che} . The initial coordinates of these simulations were generated by randomly placing one solute molecule and 1000 or 300 molecules of water or cyclohexane, respectively, in cubic computational boxes of dimensions appropriate for the experimental density of the pure solvent. After energy minimization, the systems were equilibrated as described for the pure liquids (section II.3). For the free-energy calculations, the solute–solvent interactions were perturbed using a coupling-parameter λ based on a soft-core scheme,⁷⁶ where $\lambda = 0$ corresponds to full and $\lambda = 1$ to vanishing solute–solvent interactions. The electrostatic soft-core parameter was set to 0.5 nm and the Lennard-Jones soft-core parameter to 0.5. Thermodynamic integration⁷⁷ was applied on the basis of 21 equidistant λ values and trapezoidal integration in a semi-sequential way; i.e., coordinates obtained after 50 ps of equilibration at λ were used as initial coordinates for the simulation at $\lambda + \Delta\lambda$. For each λ value, the system was equilibrated for 0.1 ns, followed by a production simulation of 0.8 ns. The estimates of ΔG_{wat} and ΔG_{che} calculated in this way, expressed on a per mole basis, were compared directly to experimental data according to a standard-state definition involving identical reference molar volumes, e.g., $1 \text{ dm}^3 \text{ mol}^{-1}$, for the gas-phase and solute standard states. This procedure performed well for all compounds, with the exception of DXE. For this compound, an additional stochastic thermostat was required for appropriate sampling.

III. RESULTS AND DISCUSSION

The results of the calculations involving the different parameter sets considered (Tables 3–5) are reported and compared to experimental data in Tables 6–11, ordered by chemical function. The overall quality of the parametrization for the different classes of compounds is also characterized in Table 12 in the form of root-mean-square deviations (RMSD) and average deviations (AVED) of the values of the different observables relative to the experimental values. The comparison is also illustrated graphically in Figure 2. For each class of compounds, a comparison of the results obtained using the different force fields is made. In nearly all cases, the present reparametrization results in a significant improvement. The word significant is used here with reference to differences on the order of at least 1% in terms of the density or at least $k_{\text{B}}T$ (2.5 kJ mol^{-1} at room temperature) in terms of energetic quantities.

III.1. Alcohols. The parameters for the alcohol function were calibrated on the basis of the primary alcohols ETL and PPL, and their transferability was subsequently tested using the other primary alcohols MTL, BTL, PTL, HXL, HPL, and OTL, the secondary alcohols 2PPL, 2BTL, 2PTL, 3PTL, and CHXL, and

Table 6. Comparison between Experimental and Simulated Properties of the Alcohols^a

param. set	compound	ρ_{liq} (kg m ⁻³)	ΔH_{vap} (kJ mol ⁻¹)	ΔG_{wat} (kJ mol ⁻¹)	ΔG_{che} (kJ mol ⁻¹)
experiment	MTL	784 ^b	37.4 ^b	-21.4 ^c	-5.4 ^c
	ETL	785 ^b	42.3 ^b	-20.9 ^c	-10.8 ^c
	PPL	800 ^b	47.5 ^b	-20.6 ^c	-11.4 ^c
	BTL	806 ^b	52.3 ^b	-19.8 ^c	-14.7 ^c
	PTL	811 ^b	56.9 ^b	-18.7 ^c	-15.1 ^c
	HXL	815 ^b	61.9 ^b	-18.2 ^c	-22.2 ^c
	HPL	820 ^d	66.8 ^c	-17.7 ^c	-25.2 ^c
	OTL	822 ^b	71.0 ^b	-17.1 ^c	
	2PPL	780 ^b	45.5 ^b	-19.3 ^c	-9.9 ^c
	2BTL	802 ^b	49.6 ^b	-19.2 ^f	
	2PTL	805 ^b	53.1 ^b	-18.4 ^f	
	3PTL	816 ^b	53.1 ^b	-18.2 ^f	
	CHXL	968 ^b	62.0 ^b	-22.9 ^f	
	2M2P	781 ^b	46.8 ^b	-18.9 ^c	-12.3 ^c
	2M2B	805 ^b	50.2 ^b	-18.5 ^f	
	53A5	MTL	811 [1.03]	41.7 {4.3}	-20.1 ± 1.0 {1.2}
ETL		778 [0.99]	45.4 {3.1}	-15.5 ± 0.9 {5.4}	-8.2 ± 0.7 {2.7}
PPL		787 [0.98]	49.8 {2.3}	-15.6 ± 1.2 {5.0}	-11.9 ± 0.9 {-0.5}
BTL		796 [0.99]	54.7 {2.4}	-14.4 ± 1.7 {5.3}	-15.7 ± 1.1 {-1.0}
PTL		803 [0.99]	59.9 {3.0}	-13.4 ± 1.5 {5.3}	-19.0 ± 1.0 {-3.9}
HXL		808 [0.99]	65.2 {3.3}	-12.2 ± 1.4 {6.1}	-23.1 ± 1.1 {-0.9}
HPL		811 [0.99]	70.1 {3.3}	-12.5 ± 1.4 {5.2}	-26.1 ± 1.2 {-0.9}
OTL		815 [0.99]	75.2 {4.2}	-12.0 ± 1.7 {5.1}	-28.4 ± 1.1 {-}
2PPL		738 [0.95]	44.3 {-1.2}	-10.6 ± 1.3 {8.7}	-9.2 ± 0.8 {0.7}
2BTL		769 [0.96]	49.3 {-0.4}	-10.3 ± 1.4 {8.9}	-13.7 ± 0.9 {-}
2PTL		779 [0.97]	54.2 {1.1}	-9.8 ± 1.2 {8.5}	-18.0 ± 0.9 {-}
3PTL		784 [0.96]	53.9 {0.8}	-10.1 ± 1.4 {8.1}	-18.4 ± 1.2 {-}
CHXL		903 [0.93]	60.4 {-1.6}	-13.2 ± 1.3 {9.7}	-23.2 ± 1.0 {-}
2M2P		735 [0.94]	39.4 {-7.4}	-7.7 ± 1.5 {11.2}	-10.4 ± 0.9 {1.9}
2M2B		772 [0.96]	45.4 {-4.8}	-7.3 ± 1.3 {11.2}	-14.6 ± 1.1 {-}
53A6		MTL	859 [1.10]	47.0 {9.6}	-22.2 ± 1.2 {-0.8}
	ETL	796 [1.01]	49.7 {7.4}	-19.3 ± 1.2 {1.6}	-8.1 ± 0.7 {2.7}
	PPL	797 [1.00]	53.9 {6.4}	-19.0 ± 1.5 {1.5}	-11.9 ± 0.9 {-0.5}
	BTL	804 [1.00]	58.9 {6.6}	-18.0 ± 1.3 {1.7}	-15.7 ± 1.1 {-0.9}
	PTL	808 [1.00]	64.0 {7.1}	-17.8 ± 1.5 {0.9}	-19.0 ± 1.0 {-3.9}
	HXL	811 [1.00]	68.9 {7.0}	-16.6 ± 1.7 {1.6}	-23.1 ± 1.1 {-0.9}
	HPL	815 [0.99]	73.1 {6.3}	-16.4 ± 1.6 {1.3}	-26.0 ± 1.2 {-0.9}
	OTL	817 [0.99]	78.0 {7.0}	-15.5 ± 1.5 {1.6}	-28.4 ± 1.1 {-}
	2PPL	745 [0.96]	48.2 {2.7}	-14.8 ± 1.3 {4.5}	-9.2 ± 0.8 {0.7}
	2BTL	773 [0.96]	52.8 {3.1}	-14.8 ± 1.4 {4.3}	-13.7 ± 0.9 {-}
	2PTL	786 [0.98]	58.4 {5.3}	-14.0 ± 1.2 {4.4}	-18.0 ± 0.9 {-}
	3PTL	789 [0.97]	57.8 {4.7}	-14.2 ± 1.5 {4.0}	-18.4 ± 1.2 {-}
	CHXL	909 [0.94]	64.2 {2.2}	-17.3 ± 1.5 {5.6}	-23.2 ± 1.0 {-}
	2M2P	742 [0.95]	43.6 {-3.2}	-12.5 ± 1.5 {6.4}	-10.4 ± 1.0 {1.9}
	2M2B	781 [0.97]	49.9 {-0.3}	-12.4 ± 1.1 {6.1}	-14.6 ± 1.1 {-}
	53A6 _{OXY}	MTL	822.7 [1.05]	43.0 {5.6}	-21.7 ± 1.4 {-0.3}
ETL		776.4 [0.99]	45.6 {3.3}	-19.5 ± 1.2 {1.5}	-6.2 ± 0.9 {4.7}
PPL		781.0 [0.98]	49.5 {2.0}	-18.2 ± 1.7 {2.4}	-10.4 ± 1.1 {1.0}
BTL		791.4 [0.98]	54.4 {2.1}	-16.5 ± 1.8 {3.3}	-14.4 ± 1.1 {0.3}
PTL		796.4 [0.98]	59.7 {2.8}	-16.1 ± 1.8 {2.6}	-17.9 ± 1.3 {-2.8}
HXL		803.0 [0.99]	64.6 {2.7}	-15.7 ± 1.8 {2.6}	-21.5 ± 1.5 {0.8}
HPL		807.3 [0.98]	69.7 {2.9}	-14.3 ± 1.8 {3.4}	-23.5 ± 1.3 {1.7}
OTL		810.6 [0.99]	74.6 {3.6}	-14.6 ± 2.1 {2.5}	-27.7 ± 1.3 {-}
2PPL		733.5 [0.94]	44.2 {-1.3}	-13.9 ± 1.3 {5.4}	-7.7 ± 1.1 {2.2}
2BTL		762.3 [0.95]	49.1 {-0.6}	-13.9 ± 1.4 {5.3}	-13.2 ± 1.2 {-}
2PTL		774.1 [0.96]	54.0 {0.9}	-12.7 ± 1.4 {5.7}	-16.0 ± 1.1 {-}
3PTL		779.1 [0.95]	53.5 {0.4}	-13.7 ± 1.5 {4.5}	-15.1 ± 1.4 {-}
CHXL		900.6 [0.93]	60.6 {-1.4}	-16.0 ± 1.7 {6.9}	-22.9 ± 1.6 {-}
2M2P		731.9 [0.94]	39.7 {-7.1}	-10.4 ± 1.4 {8.4}	-10.4 ± 1.1 {1.9}
2M2B		768.6 [0.95]	45.3 {-4.9}	-10.7 ± 1.3 {7.8}	-14.5 ± 1.5 {-}

^a The properties are the pure-liquid density ρ_{liq} and enthalpy of vaporization ΔH_{vap} as well as the solvation free energies in water ΔG_{wat} and in cyclohexane ΔG_{che} at 1 bar and 298.15 K. The parameter sets considered are 53A5 and 53A6, as well as 53A6_{OXY} (present work). The ρ_{liq} ratio (simulation divided by experiment) is indicated between square brackets. The ΔH_{vap} , ΔG_{wat} and ΔG_{che} deviations (simulation minus experiment) are indicated between braces. Experimental values were taken from the following sources: ^b ref 65, ^c refs 90 and 91, ^d ref 92, ^e ref 89, ^f ref 93.

Table 7. Comparison between Experimental and Simulated Properties of the Ethers^a

param. set	compound	ρ_{liq} (kg m ⁻³)	ΔH_{vap} (kJ mol ⁻¹)	ΔG_{wat} (kJ mol ⁻¹)	ΔG_{che} (kJ mol ⁻¹)
experiment	DME ^b	722 ^c	21.5 ^d	-8.0 ^e	
	DEE	708 ^f	27.1 ^f	-7.4 ^e	-12.7 ^e
	MPH	736 ^d	27.6 ^g	-7.0 ^e	
	DXE	864 ^f	36.4 ^f	-20.2 ^e	
53A5(6)	DME ^b	750 [1.04]	22.9 {1.4}	1.4 ± 1.0 {9.4}	-9.8 ± 0.8 {-}
	DEE	710 [1.00]	27.7 {0.6}	2.5 ± 1.1 {9.9}	-15.8 ± 1.0 {-3.1}
	MPH	726 [0.99]	29.1 {1.5}	2.1 ± 1.1 {9.1}	-15.4 ± 1.3 {-}
	DXE	855 [0.99]	37.4 {1.0}	-3.5 ± 1.6 {16.7}	-21.1 ± 1.2 {-}
53A6 _W	DME ^b	749 [1.04]	25.4 {3.9}	-7.4 ± 1.0 {0.7}	-10.8 ± 0.8 {-}
	DEE	717 [1.01]	30.7 {3.6}	-2.7 ± 1.1 {4.7}	-16.7 ± 1.0 {-4.0}
	MPH	730 [0.99]	31.4 {3.7}	-1.5 ± 1.3 {5.4}	-18.2 ± 1.1 {-}
	DXE	852 [0.99]	41.7 {5.3}	-15.5 ± 1.8 {4.7}	-22.9 ± 1.5 {-}
53A6 _{OXY}	DME ^b	713 [0.99]	24.0 {2.5}	-8.4 ± 1.1 {-0.4}	-7.3 ± 0.9 {-}
	DEE	691 [0.98]	28.0 {0.9}	-6.6 ± 1.6 {0.8}	-13.0 ± 1.4 {-0.3}
	MPH	704 [0.96]	29.3 {1.7}	-6.1 ± 1.5 {0.9}	-13.5 ± 1.2 {-}
	DXE	803 [0.93]	35.7 {-0.7}	-18.5 ± 1.8 {1.7}	-16.9 ± 1.5 {-}

^aThe properties are the pure-liquid density ρ_{liq} and enthalpy of vaporization ΔH_{vap} as well as the solvation free energies in water ΔG_{wat} and in cyclohexane ΔG_{che} at 1 bar and 298.15 K. The parameter sets considered are 53A5(6) and the variant 53A6_W, as well as 53A6_{OXY} (present work). The ρ_{liq} ratio (simulation divided by experiment) is indicated between square brackets. The ΔH_{vap} , ΔG_{wat} and ΔG_{che} deviations (simulation minus experiment) are indicated between braces. ^bAt 254 K. Experimental values were taken from the following sources: ^cref 94, ^dref 92, ^eref 90 and 91, ^fref 65, ^gref 89.

the tertiary alcohols 2M2P and 2M2B. The calculated properties are compared to experimental data in Table 6.

Considering the primary alcohols beyond MTL, the three sets (53A5, 53A6, and 53A6_{OXY}) reproduce well the experimental ρ_{liq} and ΔG_{che} . Although parameters appropriate for the longer-chain aliphatic alcohols do not appear to be transferable to MTL, no additional effort was invested in this compound considering the availability of a special parameter set for MTL within GROMOS.⁷⁸ For the secondary and tertiary alcohols, the agreement with experimental results in terms of ΔG_{che} (only two experimental values available) is also good. However, the liquid densities of these compounds are underestimated (by about 2–7%) for the three parameter sets.

Considering all alcohols, the calculated ΔH_{vap} and ΔG_{wat} differ significantly between the 53A5 and 53A6 parameter sets. On the one hand, the experimental ΔH_{vap} is reasonably well reproduced by 53A5 (RMSD of 3.4 kJ mol⁻¹), but this quantity is overestimated by 53A6 (RMSD of 5.8 kJ mol⁻¹). On the other hand, the experimental ΔG_{wat} is reasonably well reproduced by 53A6 (RMSD of 3.7 kJ mol⁻¹) but underestimated in magnitude by 53A5 (RMSD of 7.5 kJ mol⁻¹). These observations are not surprising considering that the 53A5 set was calibrated to reproduce primarily pure-liquid properties and the 53A6 set calibrated to reproduce primarily solvation properties. However, as evidenced by the results obtained using the 53A6_{OXY} parameter set, these two requirements are actually not entirely incompatible, provided that the C₆ parameter is also included in the optimization procedure. The latter set is as accurate as 53A5 in terms of ΔH_{vap} (RMSD 3.4→3.3 kJ mol⁻¹) and only slightly less accurate than 53A6 in terms of ΔG_{wat} (RMSD 3.7→4.8 kJ mol⁻¹).

Considering primary, secondary, and tertiary alcohols separately, the deviations from experimental results in terms of ΔH_{vap} and ΔG_{wat} are largely systematic within each of the three parameter sets. ΔH_{vap} is typically overestimated for primary and, to a lesser extent, secondary alcohols, while this quantity is systematically underestimated for tertiary alcohols.

III.2. Ethers. The parameters for the ether function in 53A6_{OXY} were calibrated on the basis of the topologically symmetric ethers DME and DEE, and their transferability subsequently tested using the asymmetric ether MPH and the diether DXE. Note that the parameter sets 53A5 and 53A6 are identical for these compounds and were only optimized against pure-liquid properties, because no ether group is found in amino acid side chains. In this case, the set will be referred to as 53A5(6). Note also that, in 53A6_{OXY}, the Lennard-Jones interaction parameters of the atom type OE (ether or ester oxygen) were calibrated exclusively considering esters (section III.6), and not ethers. The parameter refinement for the latter compounds thus exclusively involved the partial charges. The calculated properties are compared to experimental data in Table 7.

The 53A5(6) parameter set reproduces well the liquid properties ρ_{liq} and ΔH_{vap} , as well as ΔG_{che} for which only one experimental value is available, but it is quite inaccurate in terms of ΔG_{wat} . Here again, the 53A6_{OXY} parameter set is marginally less accurate than 53A5(6) in terms of ΔH_{vap} (RMSD 1.2→1.6 kJ mol⁻¹), while it achieves a significantly improved agreement with experimental values in terms of ΔG_{wat} (RMSD 11.7→1.1 kJ mol⁻¹). Work is currently in progress to further improve the description of diethers, in which a recalibration of the torsional potential associated with the O–C–C–O dihedral angle might be required to reproduce the experimental relative stability of different conformers.⁷⁹ Note finally that the 53A6_W variant³⁷ represented a significant improvement over 53A5(6) in terms of ΔG_{wat} . However, it is still slightly less accurate compared to 53A6_{OXY} in terms of this property, and shows a more significant deviation for ΔH_{vap} .

III.3. Aldehydes. The parameters for the aldehyde function in 53A6_{OXY} were calibrated on the basis of the compound EAL, and their transferability was subsequently tested using the compounds PAL and BAL. The hydrogen atom type HC, originally introduced for aromatic ring hydrogen atoms and characterized by nonzero Lennard-Jones interaction parameters, was selected

Table 8. Comparison between Experimental and Simulated Properties of the Aldehydes^a

param. set	compound	ρ_{liq} (kg m ⁻³)	ΔH_{vap} (kJ mol ⁻¹)	ΔG_{wat} (kJ mol ⁻¹)	ΔG_{che} (kJ mol ⁻¹)	
experiment	EAL	778 ^b	26.1 ^b	-14.6 ^c	[-]	
	PAL	791 ^b	29.6 ^b	-14.4 ^c	[-]	
	BAL	796 ^b	33.6 ^b	-13.3 ^c	[-]	
53A6 _{OXY}	EAL	789	[1.01] 30.5	{ 4.4} -13.3 ± 1.0	{ 1.3} -7.2 ± 0.7	{ -}
	PAL	779	[0.99] 34.1	{ 4.5} -11.4 ± 1.2	{ 3.0} -10.6 ± 0.8	{ -}
	BAL	790	[0.99] 38.5	{ 4.8} -11.3 ± 1.4	{ 2.0} -13.3 ± 1.1	{ -}

^aThe properties are the pure-liquid density ρ_{liq} and enthalpy of vaporization ΔH_{vap} as well as the solvation free energies in water ΔG_{wat} and in cyclohexane ΔG_{che} at 1 bar and 298.15 K. The parameter set considered is 53A6_{OXY} (present work). The ρ_{liq} ratio (simulation divided by experiment) is indicated between square brackets. The ΔH_{vap} , ΔG_{wat} , and ΔG_{che} deviations (simulation minus experiment) are indicated between braces. Experimental values were taken from the following sources: ^b ref 65, ^c ref 93.

Table 9. Comparison between Experimental and Simulated Properties of the Ketones^a

param. set	compound	ρ_{liq} (kg m ⁻³)	ΔH_{vap} (kJ mol ⁻¹)	ΔG_{wat} (kJ mol ⁻¹)	ΔG_{che} (kJ mol ⁻¹)		
experiment	PPN	784 ^b	31.3 ^b	-16.1 ^c	-11.2 ^c		
	BTN	800 ^b	34.5 ^b	-15.7 ^c	-14.6 ^c		
	2PN	802 ^b	38.4 ^b	-14.8 ^c	-17.5 ^c		
	3PN	809 ^b	38.5 ^b	-14.3 ^c	-18.0 ^c		
	2HN	807 ^b	42.9 ^b	-13.8 ^c	-20.0 ^c		
	3HN	815 ^d	42.5 ^d				
	53A5(6)	PPN	824	[1.05] 31.6	{ 0.3} -8.2 ± 0.9	{ 7.9} -13.6 ± 0.9	{ -2.5}
BTN		810	[1.01] 34.6	{ 0.1} -6.3 ± 1.0	{ 9.4} -16.3 ± 0.9	{ -1.8}	
2PN		814	[1.01] 39.5	{ 1.1} -4.7 ± 1.3	{ 10.1} -20.8 ± 1.0	{ -3.3}	
3PN		804	[0.99] 37.8	{ -0.7} -4.9 ± 1.1	{ 9.4} -19.4 ± 1.0	{ -1.4}	
2HN		817	[1.01] 44.1	{ 1.2} -5.4 ± 1.4	{ 8.3} -24.7 ± 1.1	{ -4.7}	
3HN		807	[0.99] 42.4	{ -0.1} -1.8 ± 1.5	{ -}	-22.4 ± 1.0	{ -}
53A6 _{OXY}		PPN	813	[1.04] 32.9	{ 1.6} -13.8 ± 1.2	{ 2.3} -10.8 ± 0.8	{ 0.4}
	BTN	799	[1.00] 35.5	{ 1.0} -13.0 ± 1.5	{ 2.7} -13.9 ± 1.1	{ 0.7}	
	2PN	804	[1.00] 40.6	{ 2.2} -11.8 ± 1.6	{ 3.0} -18.7 ± 0.9	{ -1.2}	
	3PN	794	[0.98] 38.7	{ 0.1} -11.0 ± 1.4	{ 3.3} -17.8 ± 1.1	{ 0.2}	
	2HN	809	[1.00] 45.2	{ 2.3} -12.1 ± 1.4	{ 1.7} -23.4 ± 1.1	{ -3.4}	
	3HN	799	[0.98] 43.3	{ 0.8} -9.9 ± 1.5	{ -}	-19.8 ± 1.3	{ -}

^aThe properties are the pure-liquid density ρ_{liq} and enthalpy of vaporization ΔH_{vap} as well as the solvation free energies in water ΔG_{wat} and in cyclohexane ΔG_{che} at 1 bar and 298.15 K. The parameter sets considered are 53A5(6) as well as 53A6_{OXY} (present work). The ρ_{liq} ratio (simulation divided by experiment) is indicated between square brackets. The ΔH_{vap} , ΔG_{wat} , and ΔG_{che} deviations (simulation minus experiment) are indicated between braces. Experimental values were taken from the following sources: ^b ref 65, ^c ref 90 and 91, ^d ref 89.

to represent the hydrogen atom of this functional group. Note that the Lennard-Jones interaction parameters of the atom type O (carbonyl oxygen) were calibrated exclusively considering ketones (section III.4), and not aldehydes. The parameter refinement for the latter compounds thus exclusively involved the partial charges. The calculated properties are compared to experimental data in Table 8. The experimental ρ_{liq} and ΔG_{wat} are very well reproduced, while ΔH_{vap} is systematically overestimated by about 4.5 kJ mol⁻¹.

III.4. Ketones. The parameters for the ketone function in 53A6_{OXY} were calibrated on the basis of the compounds PPN and BTN, and their transferability was subsequently tested using the compounds 2PN, 3PN, 2HN, and 3HN. Note that the parameter sets 53A5 and 53A6 are identical for these compounds and were only optimized against pure-liquid properties, because no ketone group is found in the amino acid side chains. In this case, the set will be referred to as 53A5(6). The calculated properties are compared to experimental data in Table 9.

Here again, the 53A6_{OXY} parameter set reproduces the ρ_{liq} and ΔG_{che} very well, slightly better than 53A5(6), and is only marginally less accurate than 53A5(6) in terms of ΔH_{vap} (RMSD 0.7→1.5 kJ mol⁻¹), while it achieves a significantly improved agreement with experimental values in terms of ΔG_{wat} (RMSD 9.1→2.7 kJ mol⁻¹).

III.5. Carboxylic Acids. The parameters for the carboxylic acid function in 53A6_{OXY} were calibrated on the basis of the compound ACA, and their transferability was subsequently tested using the compounds PPA and BTA. Note that the Lennard-Jones interaction parameters of the atom types O (carbonyl oxygen) and OA (alcohol or carboxylic acid oxygen) were calibrated exclusively considering ketones and alcohols, respectively (sections III.4 and III.1), and not carboxylic acids. The parameter refinement for the latter compounds thus exclusively involved the partial charges. The calculated properties are compared to experimental data in Table 10.

Similarly to the situation encountered for alcohols (section III.1), the experimental ΔH_{vap} is reasonably well reproduced by

Table 10. Comparison between Experimental and Simulated Properties of the Carboxylic Acids^a

param. set	compound	ρ_{liq} (kg m ⁻³)	ΔH_{vap} (kJ mol ⁻¹)	ΔG_{wat} (kJ mol ⁻¹)	ΔG_{che} (kJ mol ⁻¹)
experiment	ACA	1044 ^b	51.6 ^c	-28.3 ^d	-7.2 ^d
	PPA	988 ^b	58.6 ^c	-27.1 ^d	-15.8 ^d
	BTA	953 ^b	57.3 ^c	-26.6 ^d	
53A5	ACA	1148 [1.10]	48.0 { -3.6 }	-17.2 ± 1.1 { 11.1 }	-14.2 ± 0.9 { -7.0 }
	PPA	1034 [1.05]	50.6 { -6.7 }	-16.0 ± 1.4 { 11.1 }	-17.4 ± 1.0 { -1.6 }
	BTA	992 [1.04]	55.3 { -2.7 }	-13.4 ± 1.5 { 13.2 }	-20.9 ± 1.1 { - }
53A6	ACA	1147 [1.10]	64.2 { 12.6 }	-28.9 ± 1.3 { -0.6 }	-14.9 ± 1.0 { -7.7 }
	PPA	1029 [1.04]	67.8 { 10.5 }	-28.4 ± 1.3 { -1.3 }	-17.6 ± 1.0 { -1.8 }
	BTA	990 [1.04]	72.7 { 14.7 }	-27.3 ± 1.8 { -0.7 }	-20.5 ± 0.9 { - }
53A6 _{OXY}	ACA	1077.7 [1.03]	57.8 { 6.2 }	-24.3 ± 1.4 { 4.0 }	-9.9 ± 1.3 { -2.7 }
	PPA	975.6 [0.99]	61.3 { 4.0 }	-22.7 ± 1.5 { 4.4 }	-14.9 ± 1.3 { 0.9 }
	BTA	950.0 [1.00]	66.5 { 8.5 }	-21.4 ± 1.8 { 5.2 }	-17.9 ± 1.2 { - }

^aThe properties are the pure-liquid density ρ_{liq} and enthalpy of vaporization ΔH_{vap} as well as the solvation free energies in water ΔG_{wat} and in cyclohexane ΔG_{che} at 1 bar and 298.15 K. The parameter sets considered are 53A5 and 53A6, as well as 53A6_{OXY} (present work). The ρ_{liq} ratio (simulation divided by experiment) is indicated between square brackets. The ΔH_{vap} , ΔG_{wat} and ΔG_{che} deviations (simulation minus experiment) are indicated between braces. Experimental values were taken from the following sources: ^b ref 65, ^c ref 95, ^d refs 90 and 91.

Table 11. Comparison between Experimental and Simulated Properties of the Esters^a

param. set	compound	ρ_{liq} (kg m ⁻³)	ΔH_{vap} (kJ mol ⁻¹)	ΔG_{wat} (kJ mol ⁻¹)	ΔG_{che} (kJ mol ⁻¹)
experiment	EAE	895 ^b	35.6 ^b	-13.1 ^c	-14.9 ^c
	MPE	915 ^d	35.6 ^c	-12.3 ^c	-15.5 ^c
	PAE	883 ^b	39.8 ^b	-12.0 ^c	-18.2 ^c
	BAE	876 ^b	43.6 ^b	-10.7 ^c	-20.7 ^c
53A5	EAE	923 [1.03]	38.0 { 2.4 }	0.8 ± 1.5 { 13.9 }	-22.4 ± 1.1 { -7.5 }
	MPE	926 [1.01]	37.3 { 1.7 }	1.0 ± 1.2 { 13.3 }	-22.1 ± 1.1 { -6.6 }
	PAE	907 [1.03]	42.7 { 2.9 }	2.6 ± 1.5 { 14.5 }	-23.9 ± 1.0 { -5.7 }
	BAE	896 [1.02]	47.9 { 4.3 }	3.5 ± 1.5 { 14.1 }	-29.3 ± 1.1 { -8.7 }
53A6	EAE	959 [1.07]	47.6 { 12.0 }	-16.4 ± 1.2 { -3.4 }	-21.3 ± 1.0 { -6.4 }
	MPE	965 [1.05]	48.8 { 13.2 }	-17.3 ± 1.2 { -5.1 }	-21.4 ± 0.9 { -5.8 }
	PAE	934 [1.06]	51.5 { 11.7 }	-15.7 ± 1.4 { -3.7 }	-25.1 ± 0.9 { -6.9 }
	BAE	918 [1.05]	56.0 { 12.4 }	-15.0 ± 1.5 { -4.3 }	-26.6 ± 1.2 { -6.0 }
53A6 _{OXY}	EAE	881.8 [0.99]	36.0 { 0.4 }	-9.6 ± 1.2 { 3.4 }	-17.4 ± 1.4 { -2.5 }
	MPE	883.1 [0.97]	36.0 { 0.4 }	-10.2 ± 1.4 { 2.1 }	-16.5 ± 1.3 { -1.0 }
	PAE	871.4 [0.99]	40.4 { 0.6 }	-10.3 ± 1.7 { 1.7 }	-19.9 ± 1.3 { -1.7 }
	BAE	866.1 [0.99]	45.3 { 1.7 }	-7.4 ± 1.9 { 3.2 }	-23.3 ± 1.5 { -2.7 }

^aThe properties are the pure-liquid density ρ_{liq} and enthalpy of vaporization ΔH_{vap} as well as the solvation free energies in water ΔG_{wat} and in cyclohexane ΔG_{che} at 1 bar and 298.15 K. The parameter sets considered are 53A5 and 53A6, as well as 53A6_{OXY} (present work). The ρ_{liq} ratio (simulation divided by experiment) is indicated between square brackets. The ΔH_{vap} , ΔG_{wat} and ΔG_{che} deviations (simulation minus experiment) are indicated between braces. Experimental values were taken from the following sources: ^b ref 65, ^c ref 90 and 91, ^d ref 92, ^e ref 89.

53A5 (RMSD of 5.2 kJ mol⁻¹), this quantity being overestimated by 53A6 (RMSD of 12.7 kJ mol⁻¹), while the experimental ΔG_{wat} is very well reproduced by 53A6 (RMSD of 0.9 kJ mol⁻¹), this quantity being underestimated in magnitude by 53A5 (RMSD of 11.8 kJ mol⁻¹). In contrast to the alcohols, however, neither of the two sets performs very well in terms of ρ_{liq} (overestimated by 4–10%) and ΔG_{che} (RMSD of 5.1–5.6 kJ mol⁻¹). The 53A6_{OXY} parameter set provides a compromise between the two sets, being somewhat less accurate than 53A5 in terms of ΔH_{vap} (RMSD 5.2→6.6 kJ mol⁻¹) and noticeably less accurate than 53A6 in terms of ΔG_{wat} (RMSD 0.9→4.6 kJ mol⁻¹). However, it represents a clear improvement over both sets in terms of ρ_{liq} (maximal deviation of 3%) and ΔG_{che} (RMSD 5.1–5.6→2.0 kJ mol⁻¹).

III.6. Esters. The parameters for the ester function in 53A6_{OXY} were calibrated on the basis of the compounds EAE and PAE, and

their transferability was subsequently tested using the compounds BAE and MPE. Note that the Lennard-Jones interaction parameters of the atom type O (carbonyl oxygen) were calibrated exclusively considering ketones (section III.4), and not esters. The parameter refinement for the latter compounds thus exclusively involved the atom type OE (ether or ester oxygen) and the partial charges. The calculated properties are compared to experimental data in Table 11.

In the context of esters, 53A5 does not perform very well. It systematically overestimates ρ_{liq} (by about 3%), ΔH_{vap} (RMSD of 3.0 kJ mol⁻¹), and the magnitude of ΔG_{che} (RMSD of 7.2 kJ mol⁻¹) and, most importantly, largely underestimates the magnitude of ΔG_{wat} (RMSD of 14.0 kJ mol⁻¹). The 53A6 set³³ provides a significant improvement in terms of solvation properties ΔG_{wat} and, to a lesser extent, ΔG_{che} , but at the cost of

Table 12. Root-Mean-Square Deviations (RMSD) of the Simulation Averaged Values of Different Observables (ρ_{liq} , ΔH_{vap} , ΔG_{wat} , and ΔG_{che}) Relative to Experimental Values^a

chemical function	parameter set	ρ_{liq}		ΔH_{vap}		ΔG_{wat}		ΔG_{che}	
		(kg m ⁻³)	(kg m ⁻³)	(kJ mol ⁻¹)	(kJ mol ⁻¹)	(kJ mol ⁻¹)	(kJ mol ⁻¹)	(kJ mol ⁻¹)	(kJ mol ⁻¹)
alcohols	S3A5	29.7	[−20.7]	3.4	[0.8]	7.5	[7.0]	1.8	[−0.1]
	S3A6	31.2	[−11.2]	5.8	[4.8]	3.7	[3.0]	1.9	[−0.1]
	S3A6 _{OXY}	33.8	[−24.1]	3.3	[0.7]	4.8	[4.1]	2.2	[1.3]
ethers	S3A5(6)	15.6	[2.8]	1.2	[1.1]	11.7	[11.3]	3.1	[3.1]
	S3A6 _W	15.7	[4.5]	4.2	[4.1]	4.3	[3.9]	4.0	[4.0]
	S3A6 _{OXY}	35.8	[−29.8]	1.6	[1.1]	1.1	[0.8]	0.3	[0.3]
aldehydes	S3A6 _{OXY}	10.0	[−2.3]	4.6	[4.6]	2.2	[2.1]		
ketones	S3A5(6)	18.4	[9.8]	0.7	[0.3]	9.1	[9.0]	3.0	[−2.7]
	S3A6 _{OXY}	14.9	[0.2]	1.5	[1.4]	2.7	[2.6]	1.7	[−0.7]
carboxylic acids	S3A5	69.4	[63.0]	5.2	[−4.5]	11.8	[11.8]	5.1	[−4.3]
	S3A6	66.8	[60.3]	12.7	[12.4]	0.9	[−0.9]	5.6	[−4.8]
	S3A6 _{OXY}	20.6	[5.7]	6.6	[6.0]	4.6	[4.5]	2.0	[−0.9]
esters	S3A5	21.7	[20.8]	3.0	[2.8]	14.0	[14.0]	7.2	[−7.1]
	S3A6	52.3	[51.8]	12.3	[12.3]	4.1	[−4.1]	6.3	[−6.3]
	S3A6 _{OXY}	19.1	[−17.0]	0.9	[0.8]	2.8	[2.6]	2.1	[−1.9]
all	S3A5	30.9	[15.1]	2.7	[0.2]	10.8	[10.6]	4.0	[−3.5]
	S3A6	36.9	[22.7]	6.6	[6.2]	5.9	[3.7]	4.0	[−3.4]
	S3A6 _{OXY} ^b	24.8	[−13.0]	2.8	[2.0]	3.2	[2.9]	1.7	[−0.4]
	S3A6 _{OXY} ^c	22.4	[−11.2]	3.1	[2.5]	3.0	[2.8]	1.7	[−0.4]

^aThe corresponding average deviations (AVED) are also indicated within brackets, corresponding to simulation minus experiment. ^bExcluding the aldehydes from the calculation. ^cIncluding the aldehydes in the calculation.

deteriorating the reproduction of the pure-liquid properties ρ_{liq} and ΔH_{vap} . The S3A6_{OXY} parameter set represents an improvement over S3A5 (as well as S3A6) in terms of all properties considered, namely, ρ_{liq} (underestimated by about 1 to a maximum of 3%), ΔH_{vap} (RMSD 3.0→0.9 kJ mol⁻¹), ΔG_{wat} (RMSD 14.0→2.8 kJ mol⁻¹), and ΔG_{che} (RMSD 7.2→2.1 kJ mol⁻¹). This improvement is expected to be particularly beneficial in the context of lipid simulations, where the S3A5 parameter set did not lead to entirely satisfactory results,^{33,38,53,80–82} which may also be related to the choice of Lennard-Jones interaction parameters between headgroup atoms.³⁸

III.7. Discussion. It has sometimes been suggested that a simultaneous reproduction, with a reasonable accuracy, of the pure-liquid properties of organic liquids (molecule in a low-polarity environment) and of their hydration free energies (molecule in a high-polarity environment) is incompatible with a mean-field representation of electronic polarization effects in classical force fields.^{83,84} According to the results presented here, this statement seems to lack general validity. Furthermore, if such an incompatibility is indeed observed for some classes of compounds, it may not originate solely from electronic polarization effects but could be related to the representation of hydrogen-bonding interactions.

The ability of the S3A6_{OXY} parameter set to reproduce both pure-liquid and hydration properties differs significantly among the functional groups considered. For example, for ketones, ethers, and esters, both types of properties are reproduced accurately. In particular, deviations on the order of 1–2 kJ mol⁻¹ compared to experimental values for the hydration free

energies are within the statistical errors affecting these types of calculations and, for some compounds, within the errors associated with the experimental data itself. Therefore, for these compounds and considering the target properties ρ_{liq} , ΔH_{vap} , ΔG_{wat} , and ΔG_{che} , the present force field gives entirely satisfactory results, and an explicit representation of electronic polarizability does not appear to be required in this case. Of course, this explicit representation could still be beneficial for the description of other system properties such as dielectric or transport properties. Considering alcohols and carboxylic acids, however, although an improvement with respect to the previous versions of the force field was achieved, agreement with the experimental data is not as satisfactory. Interestingly, these compounds are also those presenting hydrogen-bond donor groups and are therefore capable of forming intermolecular hydrogen bonds in the pure-liquid state as well as of donating (rather than only accepting) hydrogen bonds to (from) water molecules in the hydrated state. Moreover, the argument usually invoked to justify the statement that implicitly polarizable classical models cannot reproduce pure-liquid and hydration properties simultaneously is that a model calibrated for a low-polarity environment will be “underpolarized” when placed in a higher-polarity environment, i.e., it will lack an additional effective dipole enhancement that should be caused by the polarization response to the new environment. Conversely, a model calibrated for a high-polarity environment will be “overpolarized” when placed in a lower polarity environment. This explanation appears reasonable but is incompatible with the observation that, for esters, the pure liquids of which have a relatively low dielectric permittivity (~5–6), the present

parametrization achieves agreement with experimental pure-liquid properties and hydration free-energies simultaneously, whereas for alcohol, the pure liquids of which have a relatively high dielectric permittivity (~ 25), it is apparently impossible to reach such an agreement. Finally, comparing the gas-phase static molecular polarizabilities of e.g. methylpropionate⁸⁵ ($\sim 8.53\text{--}8.79 \text{ \AA}^3$) and propionic acid⁸⁵ ($\sim 6.80\text{--}6.96 \text{ \AA}^3$), one would expect that polarizability effects are more important for an ester than for a carboxylic acid, although the present results, interpreted in terms of the above justification, would suggest the opposite conclusion. It seems therefore questionable to attribute the problem solely to the implicit treatment of electronic polarization effects. Classical force fields, irrespective of the way they account for molecular polarizability, involve many other approximations, ranging from the harmonic description of covalent interactions to the simplified and empirical treatment of both van der Waals interactions (e.g., *ad hoc* inverse-12th power form of the repulsion contribution to the Lennard-Jones interactions, neglect of dispersion effects beyond the inverse-sixth power term, *ad hoc* combination rules) and Coulombic interactions (e.g., distributed monopole approximation to the molecular charge density, cutoff and reaction-field correction). In particular, the apparent impossibility of reproducing simultaneously pure-liquid and hydration properties for alcohols and carboxylic acids rather hints toward deficiencies in the representation of hydrogen-bonding interactions as a mere resultant of Coulombic attraction and van der Waals repulsion. Of course, these differences may also have a component resulting from electronic polarizability. However, it should be kept in mind that even if the inclusion of explicit polarization turns out to remedy the problem, this is not sufficient to prove that electronic polarizability was its cause, considering that an increase in the flexibility of the force field functional form by introduction of new parameters will automatically improve agreement with experimental data after an appropriate parametrization.

IV. CONCLUSIONS

In the present work, a new parameter set (53A6_{OXY}) is developed for the GROMOS force field that combines reoptimized parameters for the oxygen functions (alcohols, ethers, aldehydes, ketones, carboxylic acids, and esters) with the current biomolecular force-field version³⁴ 53A6 for all other functions.

For the 35 oxygen compounds considered, the new 53A6_{OXY} parameter set provides a unified and satisfactory description of the two pure-liquid properties (ρ_{liq} , ΔH_{vap}) and the two solvation properties (ΔG_{wat} , ΔG_{che}) that were considered. Compared to 53A5 and 53A6, 53A6_{OXY} nearly systematically leads to comparable or improved agreement with experimental data in terms of these four quantities (the only noticeable exception being ΔG_{wat} for carboxylic acids).

The performance of the 53A6_{OXY} set relies on adding increased flexibility in the calibration task through (i) a simultaneous rather than separate refinement against experimental values of pure-liquid and solvation properties, (ii) a simultaneous rather than successive refinement of the Lennard-Jones interaction parameters and charges, and (iii) the allowance of moderate adjustments in the dispersive coefficients ($C_6^{1/2}$) of the Lennard-Jones interactions. Note that the latter adjustments are truly limited, with a change of the value $0.04756 \text{ (kJ mol}^{-1} \text{ nm}^6)^{1/2}$ for all oxygen atoms in 53A5 and 53A6 to 0.04136 , 0.04500 , and $0.04123 \text{ (kJ mol}^{-1} \text{ nm}^6)^{1/2}$ for the atom types O, OA, and OE, respectively, in 53A6_{OXY}.

Work is currently in progress following the same strategy to define improved interaction parameters for nitrogen-containing (amine and amide functions), sulfur-containing (thiol and sulfide functions), and aromatic compounds in the GROMOS force field. This extension would lead to an improved force field covering the entire range of amino acid side chains in natural polypeptides. A combination of 53A6_{OXY} with the recently reoptimized 53A6_{CARBO} parameter set for carbohydrates⁸⁶ is also planned. As with earlier versions of the GROMOS force field, the appropriateness of the new parametrizations remains to be validated in the context of biomolecular simulations.

AUTHOR INFORMATION

Corresponding Author

*Phone: +41 44 632 5503. Fax: +41 44 632 1039. E-mail: bruno@igc.phys.chem.ethz.ch (B.A.C.H.); phil@igc.phys.chem.ethz.ch (P.H.H.).

ACKNOWLEDGMENT

The authors would like to thank Chris Oostenbrink, Alex de Vries, Alan Mark, and Alpesh Malde for valuable discussions. Many thanks are due to the "Brutus Team", and in particular Dr. Olivier Byrde, for steady support with the computer cluster. Financial support from the Swiss National Science Foundation (Grants 21-121895 and 200020-121913), from the National Center of Competence in Research (NCCR) in Structural Biology, and from the European Research Council (Grant 228076) are also gratefully acknowledged.

REFERENCES

- (1) Allen, M. P.; Tildesley, D. J. *Computer simulation of liquids*; Oxford University Press: New York, 1987.
- (2) van Gunsteren, W. F.; Berendsen, H. J. C. *Angew. Chem., Int. Ed.* **1990**, *29*, 992–1023.
- (3) van Gunsteren, W. F.; Bakowies, D.; Baron, R.; Chandrasekhar, I.; Christen, M.; Daura, X.; Gee, P.; Geerke, D. P.; Glättli, A.; Hünenberger, P. H.; Kastenholz, M. A.; Oostenbrink, C.; Schenk, M.; Trzesniak, D.; van der Vegt, N. F. A.; Yu, H. B. *Angew. Chem., Int. Ed.* **2006**, *45*, 4064–4092.
- (4) Berendsen, H. J. C. *Simulating the physical world.*; Cambridge University Press: Cambridge, U.K., 2007.
- (5) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; di Nola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
- (6) van Gunsteren, W. F.; Berendsen, H. J. C. *Mol. Phys.* **1977**, *34*, 1311–1327.
- (7) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hünenberger, P. H.; Krüger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. *Biomolecular simulation: The GROMOS96 manual and user guide*; Verlag der Fachvereine: Zürich, Switzerland, 1996.
- (8) Scott, W. R. P.; Hünenberger, P. H.; Tironi, I. G.; Mark, A. E.; Billeter, S. R.; Fennen, J.; Torda, A. E.; Huber, T.; Krüger, P.; van Gunsteren, W. F. *J. Phys. Chem. A* **1999**, *103*, 3596–3607.
- (9) Engelsen, S. B.; Fabricius, J.; Rasmussen, K. *Acta Chem. Scand.* **1994**, *48*, 548–552.
- (10) Engelsen, S. B.; Fabricius, J.; Rasmussen, K. *Acta Chem. Scand.* **1994**, *48*, 553–565.
- (11) Gaedt, K.; Holtje, H. D. *J. Comput. Chem.* **1998**, *19*, 935–946.
- (12) Allinger, N. L. *J. Am. Chem. Soc.* **1989**, *111*, 8551–8566.
- (13) Allinger, N. L. *J. Am. Chem. Soc.* **1989**, *111*, 8566–8575.
- (14) Allinger, N. L. *J. Am. Chem. Soc.* **1989**, *111*, 8576–8582.
- (15) Allinger, N. L.; Chen, K.-H.; Lü, J.-H. *J. Comput. Chem.* **2003**, *24*, 1447–1472.

- (16) MacKerell, A. D.; Wiorkiewicz-kuczera, J.; Karplus, M. *J. Am. Chem. Soc.* **1995**, *117*, 11946–11975.
- (17) Mackerell, A. D., Jr.; 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.* **1998**, *B*, 102.
- (18) Feller, S. E.; MacKerell, A. D. *J. Phys. Chem. B* **2000**, *104*, 7510–7515.
- (19) Hatcher, E. R.; Guvench, O.; MacKerell, A. D., Jr. *J. Phys. Chem. B* **2009**, *113*, 12466–12476.
- (20) Weiner, P. K.; Kollman, P. A. *J. Comput. Chem.* **1981**, *2*, 287–303.
- (21) Pearlman, D. A.; Case, D. A.; Caldwell, J. D.; Ross, W. S.; Cheatham, T. E., III; DeBolt, S.; Fergusson, D.; Seibel, G.; Kollman, P. *Comput. Phys. Commun.* **1995**, *91*, 1–41.
- (22) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.
- (23) Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; Gonzalez-Outerino, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J. *J. Comput. Chem.* **2008**, *29*, 622–655.
- (24) Pranata, J.; Wierschke, S. G.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1991**, *113*, 2810–2819.
- (25) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.
- (26) Damm, W.; Frontera, A.; Tirado-Rives, J.; Jorgensen, W. *J. Comput. Chem.* **1997**, *18*, 1955–1970.
- (27) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem.* **2001**, *B*, 105.
- (28) Daura, X.; Mark, A. E.; van Gunsteren, W. F. *J. Comput. Chem.* **1998**, *19*, 535–547.
- (29) van Gunsteren, W. F.; Daura, X.; Mark, A. E. GROMOS force field. In *Encyclopedia of computational chemistry*; Schleyer, P., Ed.; John Wiley & Sons: Chichester, U.K., 1998; Vol. 2, pp 1211–1216.
- (30) Schuler, L. D.; van Gunsteren, W. F. *Mol. Simul.* **2000**, *25*, 301–319.
- (31) Schuler, L. D.; Daura, X.; van Gunsteren, W. F. *J. Comput. Chem.* **2001**, *22*, 1205–1218.
- (32) Chandrasekhar, I.; Kastenholz, M. A.; Lins, R. D.; Oostenbrink, C.; Schüller, L. D.; Tieleman, D. P.; van Gunsteren, W. F. *Eur. Biophys. J.* **2003**, *32*, 67–77.
- (33) Chandrasekhar, I.; Oostenbrink, C.; van Gunsteren, W. F. *Soft Mater.* **2004**, *2*, 27–45.
- (34) Oostenbrink, C.; Villa, A.; Mark, A. E.; van Gunsteren, W. F. *J. Comput. Chem.* **2004**, *25*, 1656–1676.
- (35) Soares, T. A.; Hünenberger, P. H.; Kastenholz, M. A.; Kräutler, V.; Lenz, T.; Lins, R. D.; Oostenbrink, C.; van Gunsteren, W. F. *J. Comput. Chem.* **2005**, *26*, 725–737.
- (36) Lins, R. D.; Hünenberger, P. H. *J. Comput. Chem.* **2005**, *26*, 1400–1412.
- (37) Winger, M.; de Vries, A. H.; van Gunsteren, W. F. *Mol. Phys.* **2009**, *107*, 1313–1321.
- (38) Poger, D.; van Gunsteren, W. F.; Mark, A. E. *J. Comput. Chem.* **2009**, *31*, 1117–1125.
- (39) Gee, P. J.; van Gunsteren, W. F. *Mol. Phys.* **2006**, *104*, 477–483.
- (40) Geerke, D. P.; van Gunsteren, W. F. *Mol. Phys.* **2007**, *105*, 1861–1881.
- (41) Walser, R.; Hünenberger, P. H.; van Gunsteren, W. F. *Proteins: Struct. Funct. Genet.* **2001**, *44*, 509–519.
- (42) Walser, R.; Hünenberger, P. H.; van Gunsteren, W. F. *Proteins: Struct. Funct. Genet.* **2002**, *48*, 327–340.
- (43) Smith, L. J.; Berendsen, H. J. C.; van Gunsteren, W. F. *J. Phys. Chem. B* **2004**, *108*, 1065–1071.
- (44) Geerke, D. P.; van Gunsteren, W. F. *ChemPhysChem* **2006**, *7*, 671–678.
- (45) Zagrovic, B.; Gattin, Z.; Kai-Chi Lau, J.; Huber, M.; van Gunsteren, W. F. *Eur. Biophys. J.* **2008**, *37*, 903–912.
- (46) Meier, K.; van Gunsteren, W. F. *J. Phys. Chem. A* **2010**, *114*, 1852–1859.
- (47) Allison, J. R.; van Gunsteren, W. F. *ChemPhysChem* **2009**, *10*, 3213–3228.
- (48) Eichenberger, A. P.; Gattin, Z.; Yalak, G.; van Gunsteren, W. F. *Helv. Chim. Acta* **2010**, *93*, 1857–1869.
- (49) Dolenc, J.; Oostenbrink, C.; Koller, J.; van Gunsteren, W. F. *Nucleic Acids Res.* **2005**, *33*, 725–733.
- (50) Perić-Hassler, L.; Hansen, H. S.; Baron, R.; Hünenberger, P. H. *Carbohydr. Res.* **2010**, *345*, 1781–1801.
- (51) Hansen, H. S.; Hünenberger, P. H. *J. Comput. Chem.* **2010**, *31*, 1–23.
- (52) Siwko, M. E.; de Vries, A. H.; Mark, A. E.; Kozubek, A.; Marrink, S. J. *Biophys. J.* **2009**, *96*, 3140–3153.
- (53) Horta, B. A. C.; Perić-Hassler, L.; Hünenberger, P. H. *J. Mol. Graphics Modell.* **2010**, *29*, 331–346.
- (54) Geerke, D. P.; van Gunsteren, W. F. *J. Chem. Theory Comput.* **2007**, *3*, 2128–2137.
- (55) Geerke, D. P.; van Gunsteren, W. F. *J. Phys. Chem. B* **2007**, *111*, 6425–6436.
- (56) Kunz, A.-P. E.; van Gunsteren, W. F. *J. Phys. Chem. A* **2009**, *113*, 11570–11579.
- (57) Lin, Z.; Kunz, A.-P.; van Gunsteren, W. F. *Mol. Phys.* **2010**, *108*, 1749–1757.
- (58) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. Interaction models for water in relation to protein hydration. In *Intermolecular Forces*; Pullman, B., Ed.; Reidel: Dordrecht, The Netherlands, 1981; pp 331–342.
- (59) Tironi, I. G.; Sperb, R.; Smith, P. E.; van Gunsteren, W. F. *J. Chem. Phys.* **1995**, *102*, 5451–5459.
- (60) Hermans, J.; Berendsen, H. J. C.; van Gunsteren, W. F.; Postma, J. P. M. *Biopolymers* **1984**, *23*, 1513–1518.
- (61) Daura, X.; Oliva, B.; Querol, E.; Aviles, F. X. *Proteins: Struct. Funct. Genet.* **1996**, *25*, 89–103.
- (62) Soares, T. A.; Daura, X.; Oostenbrink, C.; Smith, L. J.; van Gunsteren, W. F. *J. Biomol. NMR* **2004**, *30*, 407–422.
- (63) Oostenbrink, C.; Soares, T. A.; van der Vegt, N. F. A.; van Gunsteren, W. F. *Eur. Biophys. J.* **2005**, *34*, 273–284.
- (64) Slater, J. C.; Kirkwood, J. G. *Phys. Rev.* **1931**, *37*, 682–697.
- (65) Riddick, J. A.; Bunger, W. B.; Sakano, T. K. *Organic solvents, physical properties and methods of purification*; John Wiley & Sons: New York, 1986.
- (66) 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.; D’Annunzio, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, Revision D.01; Gaussian, Inc.: Wallingford, CT, 2004.
- (67) Villa, A.; Mark, A. E. *J. Comput. Chem.* **2002**, *23*, 548–553.
- (68) Horta, B. A. C.; de Vries, A. H.; Hünenberger, P. H. *J. Chem. Theory Comput.* **2010**, *6*, 2488–2500.
- (69) Christen, M.; Hünenberger, P. H.; Bakowies, D.; Baron, R.; Bürgi, R.; Geerke, D. P.; Heinz, T. N.; Kastenholz, M. A.; Kräutler, V.; Oostenbrink, C.; Peter, C.; Trzesniak, D.; van Gunsteren, W. F. *J. Comput. Chem.* **2005**, *26*, 1719–1751.
- (70) Hockney, R. W. *Methods Comput. Phys.* **1970**, *9*, 136–211.
- (71) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.

- (72) Barker, J. A.; Watts, R. O. *Mol. Phys.* **1973**, *26*, 789–792.
- (73) Heinz, T. N.; van Gunsteren, W. F.; Hünenberger, P. H. *J. Chem. Phys.* **2001**, *115*, 1125–1136.
- (74) Heinz, T. N.; Hünenberger, P. H. *J. Chem. Phys.* **2005**, *123*, 034107/1–034107/19.
- (75) *IUPAC Quantities, Units and Symbols in Physical Chemistry, Green Book*, 2nd ed.; Blackwell Scientific Publications: Oxford, U. K., 1993.
- (76) Beutler, T. C.; Mark, A. E.; van Schaik, R.; Gerber, P. R.; van Gunsteren, W. F. *Chem. Phys. Lett.* **1994**, *222*, 529–539.
- (77) Kirkwood, J. G. *J. Chem. Phys.* **1935**, *3*, 300–313.
- (78) Walser, R.; Mark, A. E.; van Gunsteren, W. F.; Lauterbach, M.; Wipff, G. *J. Chem. Phys.* **2000**, *112*, 10450–10459.
- (79) Fuchs, P. F. J.; Horta, B. A. C.; Hünenberger, P. H. 2010, in preparation.
- (80) Pereira, C. S.; Lins, R. D.; Chandrasekhar, I.; Freitas, L. C. G.; Hünenberger, P. H. *Biophys. J.* **2004**, *86*, 2273–2285.
- (81) Chandrasekhar, I.; Bakowies, D.; Glättli, A.; Hünenberger, P. H.; Pereira, C.; van Gunsteren, W. F. *Mol. Simul.* **2005**, *31*, 543–548.
- (82) Pereira, C. S.; Hünenberger, P. H. *J. Phys. Chem. B* **2006**, *110*, 15572–15581.
- (83) Baker, C. M.; Lopes, P. E. M.; Zhu, X.; Roux, B.; McKerell, A. D. *J. Chem. Theory Comput.* **2010**, *6*, 1181–1198.
- (84) Zhong, Y.; Patel, S. *J. Phys. Chem. B* **2010**, *114*, 11076–11092.
- (85) Miller, K. J. *J. Am. Chem. Soc.* **1990**, *112*, 8533–8542.
- (86) Hansen, H. S.; Hünenberger, P. H. *J. Comput. Chem.* **2010**, submitted.
- (87) Börjesson, U.; Hünenberger, P. H. *J. Phys. Chem. B* **2004**, *108*, 13551–13559.
- (88) Wohlfahrt, C. Pure liquids: Data. In *Landolt-Börnstein. Numerical data and functional relationships in science and technology*; Madelung, O., Ed.; Springer: Berlin, Germany, 1991; Vol. 6, pp 5–228.
- (89) Majer, V.; Svoboda, V. *Enthalpies of vaporization of organic compounds: A critical review and data compilation*; Blackwell Scientific Publications: Oxford, U.K., 1985.
- (90) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **1998**, *102*, 3257–3271.
- (91) Li, J.; Zhu, T.; Hawkins, G. D.; Winget, P.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. *Theor. Chem. Acc.* **1999**, *103*, 9–63.
- (92) Lide, D. R. *CRC Handbook of Chemistry and Physics*, 80th ed.; CRC Press: Boca Raton, FL, 1999.
- (93) Gallicchio, E.; Zhang, L. Y.; Levy, R. M. *J. Comput. Chem.* **2001**, *23*, 517–529.
- (94) Wu, J.; Liu, Z.; Bi, S.; Meng, X. *J. Chem. Eng. Data* **2003**, *48*, 426–429.
- (95) Pedley, J. B.; Naylor, R. D.; Kirby, S. P. *Thermodynamical Data of Organic Compounds*; Chapman and Hall: London, U.K., 1986.

An Analysis of the Validity of Markov State Models for Emulating the Dynamics of Classical Molecular Systems and Ensembles

Bettina Keller, Philippe Hünenberger, and Wilfred F. van Gunsteren*

Laboratory of Physical Chemistry, Swiss Federal Institute of Technology Zürich, ETH Zürich, CH-8093 Zürich, Switzerland

ABSTRACT: Markov state models parametrized using molecular simulation data are powerful tools for the investigation of conformational changes in biomolecules and in recent years have gained increasing popularity. However, a Markov state model is an approximation to the true dynamics of the complete system. We show how Markov state models are derived from the generalized Liouville equation identifying the assumptions and approximations involved and review the mathematical properties of transition matrices. Using two model systems, a two-bit flipping model consisting of only four states, and molecular dynamics simulations of liquid butane, we subsequently assess the influence of the assumptions, for example, of the marginal degrees of freedom, used in the derivation on the validity of the Markov state model.

1. INTRODUCTION

The dynamics of large biomolecules encompasses processes of vastly different time scales. Fast processes, such as bond-angle vibrations, happen on the femtosecond time scale and are coupled to slow processes, such as large conformational rearrangements, which happen on the micro- to millisecond time scale. For example, the folding of an entire protein can easily take seconds. In a molecular dynamics (MD) simulation of a biomolecule at the atomic level, the integration step of the simulation is bound to the order of 1 fs.^{1,2} Constraining the fast processes to a fixed value, the only exception being the bond-length vibration, will distort the dynamics of the slow processes.^{3–5} Thus, a MD simulation of a folding process aims at emulating a process with a time scale of micro- to milliseconds by tracing out trajectories at a femtosecond resolution, thereby bridging time scales of 9 to 12 orders of magnitude. This together with the fact that MD simulation programs scale poorly for the parallelization to many processors makes MD simulations of biomolecular processes time-consuming. Moreover, by integrating the time evolution of each degree of freedom explicitly, the amount of detail produced by an atomic-level MD simulation is barely manageable and often far beyond what is needed for the elucidation of a particular molecular phenomenon.

The combination of MD simulation with stochastic models such as Markov state models (MSM) has the power to redress both of these shortcomings. For the construction of a MSM, the complete coordinate space of the simulated system (i.e., solute plus solvent) is split into a set of *relevant* coordinates and a set of *marginal* coordinates. In the first instance, the term “relevant coordinates” denotes any set of coordinates which is of relevance to the question which is to be investigated by the simulation. MSMs provide a means to concisely represent the dynamics of the relevant coordinates, formulated as a transition matrix with a dimension on the order of typically several hundreds to several thousands. Properties of interest, such as mean first passage times and mean life times, can be directly extracted from this transition matrix.^{6,7} The conformational equilibrium distribution emerges as the first eigenvector of the transition matrix, and metastable states can be identified by grouping the states of the Markov

process in such a manner that the metastability of the groups is maximized. This corresponds to maximizing the trace of the coarse-grained matrix.⁸

MSMs are associated with a time step, called *lag time*, which is typically on the order of pico- to nanoseconds. If the dynamics of the relevant coordinates are indeed Markovian at this lag time, then the MSM can be parametrized by a large number of short MD simulations.¹⁰ This approach does not necessarily decrease the required computer time but rather, when the short simulations are run in parallel on several computers, the time one has to wait for the results.

In a MSM, the configurations of the system are mapped onto a (typically small) set of states, and the dynamics are modeled by the transition probabilities between these states. While from a mathematical point of view, the mapping corresponds to a projection and can be done by a single operator multiplication,⁹ in any practical application, this projection is split into two consecutive steps: (i) separation of the complete coordinate space into relevant and marginal coordinates and (ii) discretization of the relevant coordinates. These models are clearly an approximation of the true dynamics. The idea is that the influence of the marginal degrees of freedom averages out over time and that one can often find a time lag τ_{Markov} for which the deviation from Markovian behavior in the relevant coordinates is small enough to be neglected. A Markov model with a time resolution of τ_{Markov} or larger may then represent a realistic model of the true dynamics. Ultimately, the quality of the Markov model, i.e., how faithfully the model reproduces the dynamics in the relevant degrees of freedom, depends on (i) the interaction of the set of marginal coordinates with the set of relevant coordinates, (ii) the precise discretization of the relevant coordinates, and (iii) the statistical errors due to finite sampling of the dynamics of the system.

Since Swope et al.^{11,12} presented the first extensive and detailed application of MSM to the analysis of molecular simulation data, MSMs of biomolecular systems have developed into a very active field of research.^{9,13–17}

Received: January 28, 2011

Published: March 10, 2011

The extraction of metastable states from a given MSM, which is equivalent to the coarse-graining of the transition matrix, has been a major issue in the discussion of the application of MSMs. Two basic approaches have been published. One maximizes the metastability of the resulting coarse-grained states using temperature annealing schemes;^{8,17} the other exploits properties of the eigenvectors of the fine-grained transition matrix to define the coarse-grained states.^{7,18,19}

Recently, methods which optimize the amount of simulation data needed for the construction of a MSM have been published. These methods either rely on enhanced sampling techniques in the simulation process^{16,20–24} or apply an adaptive sampling scheme which couples the start of new simulations to a quality estimate of the current MSM.²⁵

A large number of publications deal with the discretization of the relevant degrees of freedom. Noé et al.⁷ discretized each backbone dihedral angle along the minima of a probability distribution of this angle, thereby discretizing the conformational space according to the rotamers of the molecule. More often, however, the conformations of the molecule are mapped onto more global descriptors such as secondary structure motifs of amino acids in a peptide^{13,20} or the number of intramolecular hydrogen bonds.⁷ In 2007, Chodera et al.⁸ published an adaptive discretization scheme in RMSD space. Jensen et al.²⁶ discretized the two central dihedral angles of a tetrapeptide according to the most populated regions in their Ramachandran plots and varied the positions of the boundaries. They found that the quality of the MSM is sensitive to the exact position of the boundaries. This finding is in line with the results of Sarich et al.,⁹ who demonstrate analytically that the error caused by the discretization is determined by the precision with which the transition region is discretized. Moving the boundary away from the transition point impairs the quality of the MSM.

Several methods have been developed for the estimation of the statistical uncertainties in the eigenvalues and eigenvectors of the transition matrix and properties derived from the transition matrix.^{27–29}

To the best of our knowledge, no systematic study of the influence of the marginal coordinates on the dynamics of the relevant coordinates has been published. Conceptually, this question is close to the discussion about the influence of the bath degrees of freedom on the solute coordinates in Brownian dynamics.³⁰ However, some of the assumptions (very large number of bath degrees of freedom, all coupled with the same coupling constant to the solute degrees of freedom) clearly do not apply in the context of MSM of molecular dynamics.

This publication has two objectives: (i) a review of the mathematical concepts and assumptions which form the basis of a stochastic model of molecular dynamics and (ii) an illustration of the effect of the marginal coordinates on the dynamics of the relevant coordinates. These two parts are closely linked since the properties of the marginal coordinates determine to a large extent the quality of the Markov model.

In the first part, we demonstrate how stochastic equations of motion emerge from deterministic ones when the coordinate set is split into relevant and marginal coordinates. We list the conditions a stochastic equation of motion must fulfill in order to be Markovian. We then show how the matrix formalism of transition matrices arises from a given Markovian equation of motion. Finally, we review the mathematical properties of transition matrices and link them to physical concepts such as ergodicity and equilibrium dynamics.

In the second part, we use two model systems to study how the properties of the marginal coordinates affect the assumption that the dynamics of the relevant coordinates are Markovian. The first system consists of two bits which can flip between “0” and “1”. One bit represents the relevant coordinates, the other the marginal ones. We illustrate how the coupling strength and the relative speed of the two bits influence the quality of the Markov model. The second system consists of molecular dynamics simulations of a butane molecule (relevant coordinates) immersed in a solvent of butane molecules (marginal coordinates). The solvent is modeled on the one hand explicitly (at various temperatures and pressures) and on the other hand implicitly using stochastic dynamics (with various temperatures and friction coefficients). The influence of these parameters on the quality of the Markov model is demonstrated.

2. THEORY

We consider a *system* of N_x time-dependent variables and an *ensemble* of an infinite number of replicas of this system. The *configuration* of system n in the ensemble at time t is defined by a configuration vector $\mathbf{x}_n(t)$ containing the instantaneous values of the N_x variables of this system at time t . The dynamics of the ensemble is said to be *Markovian* if the individual systems obey equations of motion of the form

$$\dot{\mathbf{x}}_n(t) = f(\mathbf{x}_n(t), \mathbf{y}_s(t, s_n)) \quad (1)$$

where f and \mathbf{y}_s on the right-hand side are functions that are identical for all systems in the ensemble, while s_n represents a scalar value attributed to a specific system n . Equation 1 states that the change of the configuration of the n th system at time t , $\dot{\mathbf{x}}_n(t)$, only depends on its current configuration $\mathbf{x}_n(t)$ and the current values of a set of variables $\mathbf{y}_s(t, s_n)$.

The function $\mathbf{y}_s(t, s_n)$ is used to implement the difference between deterministic and stochastic ensemble dynamics. If the ensemble dynamics are deterministic, the parameter s_n can be dropped and $\mathbf{y}_s(t, s_n) = \mathbf{y}_s(t)$ is the same for all systems; i.e., all systems in the ensemble follow the same equation of motion. If the ensemble dynamics involves a set of stochastic variables, $\mathbf{y}_s(t, s_n)$ represents a particular trajectory with index s_n in this variable space, which was drawn from a stochastic process $\mathbf{Y}(t)$. In this case, the equation of motion, eq 1, differs for each system.

In computational terms, s_n can be viewed as the seed for a pseudorandom number sequence assigned to the system. In more mathematical terms, s_n can also be viewed as defining this sequence itself, e.g., in the form of the representation of this real number by an infinite string of bits. The way in which s_n , or the derived pseudorandom number sequence, is exploited by the function \mathbf{y}_s , e.g., to generate the time series of stochastic Gaussian-distributed variables, need not be specified at this point. However, it is assumed that the resulting probability distribution of the stochastic variables over all systems in the ensemble at a given configuration \mathbf{x} is time-invariant.

The key assumptions for Markovian ensemble dynamics are that the single-system dynamics fulfill the following conditions: (i) *deterministic* ($\mathbf{x}_n(t)$ is determined by the sole knowledge of $\mathbf{x}_n(0)$ and s_n), *memoryless* ($\dot{\mathbf{x}}_n(t)$, as expressed by the function f , involves no explicit dependence on $\mathbf{x}_n(t')$ with $t' < t$), and *stationary* ($\dot{\mathbf{x}}_n(t)$, as expressed by the function f , involves no explicit dependence on t); (ii) a common equation of motion (no stochastic component) or a set of stochastically distributed equations of motion for the different systems in the ensemble.

Note that the stochasticity property only becomes apparent at the level of the ensemble. From the point of view of a single system n , the dynamics are entirely deterministic, given the value of s_n assigned to this system. Note also that the system dynamics need not necessarily be continuous in time; i.e., $\dot{\mathbf{x}}_n(t)$ may involve Dirac delta functions in time.

The instantaneous *macrostate* of the ensemble is defined by the (normalized) *configurational probability distribution* $\rho(\mathbf{x}, t)$ of the individual systems at time t in the N_x -dimensional space of the system configurations, which obeys the generalized Liouville equation:

$$\dot{\rho}(\mathbf{x}, t) = \hat{\mathcal{L}}\rho(\mathbf{x}, t) \quad (2)$$

$\hat{\mathcal{L}}$ is called the generalized Liouville operator or the generator. The assumptions of Markovian dynamics imply that $\hat{\mathcal{L}}$ in eq 2 is time-independent, corresponding to equilibrium dynamics. Introducing the requirement that eq 2 be valid for any arbitrary initial configurational distribution $\rho(\mathbf{x}, 0)$ and all times (including $t = 0$), the operator $\hat{\mathcal{L}}$ is unique, and its exact form can, at least in principle, be derived from knowledge of the function $f(\mathbf{x}, \mathbf{y})$ in eq 1 and of the stochastic variable probability distribution. Different forms of the generalized Liouville equation include the Liouville equation^{31,32} (Hamiltonian dynamics), the Fokker–Planck equation^{6,32} (Langevin dynamics), or the Smoluchowski equation³² (Brownian dynamics).

By introducing an infinite set of basis functions $\phi_i(\mathbf{x})$, e.g., Dirac delta functions, covering the N_x -dimensional space of the configuration variables, the configurational probability distribution $\rho(\mathbf{x}, t)$ may be rewritten as a *configurational probability vector* $\mathbf{p}(t)$ with components $p_i(t)$, which are real, non-negative, and sum up to 1. The generalized Liouville equation, eq 2, may then be translated into an equivalent matrix equation:

$$\dot{\mathbf{p}}(t) = \mathbf{K}\mathbf{p}(t) \quad (3)$$

in which the generalized Liouville matrix \mathbf{K} is a rate matrix with off-diagonal elements $K_{ij} \geq 0$ representing the rate of transition from configuration point j to configuration point i and the diagonal element K_{jj} is equal to $-\sum_{i \neq j} K_{ij}$. Consequently, the elements of each of its columns add up to 0.

Equation 3 can be formally integrated in time over an interval τ ($\tau > 0$), referred to as a *lag-time* yielding

$$\mathbf{p}(t + \tau) = \mathbf{T}(\tau)\mathbf{p}(t) \quad (4)$$

resulting in the introduction of a corresponding *transition matrix* $\mathbf{T}(\tau)$, defined as

$$\mathbf{T}(\tau) \doteq \exp(\tau\mathbf{K}) \quad (5)$$

Equation 5 effectively introduces a time discretization of the continuous Markov process. The elements of $\mathbf{T}(\tau)$ are real and non-negative

$$T_{ij} \in \mathbb{R}, T_{ij}(\tau) \geq 0 \quad \forall i, j, \tau \quad (6)$$

and satisfy the normalization condition

$$\sum_i T_{ij}(\tau) = 1 \quad \forall j, \tau \quad (7)$$

They represent the probability of a transition from a point j to a point i in configurational space during a lag time τ . Equations 6 and 7 define a *column stochastic* matrix. From the definition of $\mathbf{T}(\tau)$, one can directly derive the Chapman–Kolmogorov equation representing the *recursivity property*

$$\mathbf{T}(\tau_1 + \tau_2) = \mathbf{T}(\tau_1)\mathbf{T}(\tau_2) = \mathbf{T}(\tau_2)\mathbf{T}(\tau_1) \quad (8)$$

When formulated as

$$\mathbf{T}(n\tau) = \mathbf{T}^n(\tau) \quad (9)$$

this relation can be used as a check of whether a process with a time discretization of τ is Markovian.^{6,14}

The matrix $\mathbf{T}(\tau)$ possesses N_d *eigenvalues*, $\lambda_\alpha(\tau)$, with associated left eigenvectors, ψ_α . The eigenvectors are formally defined within an arbitrary multiplicative factor. To make their definition unambiguous, it will be assumed that these vectors are selected such that (i) the sum of the two-norm of the elements of an eigenvector is always unity; (ii) the first nonvanishing component of an eigenvector is always real and positive. With this convention, the eigenvectors with real eigenvalues always have real components that add up to unity. Because $\mathbf{T}(\tau)$ is column stochastic, it also has the following properties:¹⁸

1. It possesses a special (real) left eigenvector (which will be given the index $\alpha = 1$), $\psi_1 = N_d^{-1}\{1, 1, \dots, 1\}$ associated with the eigenvalue $\lambda_1 = 1$. Therefore, it also possesses at least one corresponding (real) right eigenvector associated with this eigenvalue. Note that in the case of uncoupled Markov chains, i.e., if $\mathbf{T}(\tau)$ can be permuted into a block-diagonal form, the eigenvalue is degenerate, i.e., associated with more than one left and right eigenvector.¹⁸
2. Its eigenvalue spectrum has a radius of 1, i.e., $|\lambda_\alpha(\tau)| \leq 1$ for all α .

At this point, one may add a third assumption to the assumptions underlying Markovian dynamics, namely that of irreducibility. The Markovian ensemble dynamics is also *irreducible* when

$$\lim_{\tau \rightarrow \infty} \tau^{-1} \int_0^\tau dt' \mathbf{T}(t') = \mathbf{T}_{\text{sum}} > 0 \quad (10)$$

where $\mathbf{T}_{\text{sum}} > 0$ is a short-hand notation for

$$\mathbf{T}_{\text{sum}, ij} > 0 \quad \forall i, j \quad (11)$$

Irreducibility implies that any configuration has a nonvanishing probability of undergoing a transition to any other configuration considering all possible lag times. It does, however, not imply that there exists a single lag time τ' for which all possible transitions have a nonvanishing transition probability.

Irreducibility is not identical with the concept of ergodicity, but it is closely linked to it. A system is *ergodic* if the time its (sufficiently long) trajectory spends in any given configuration is proportional to the probability with which this configuration is realized in the ensemble at a given time t . Then, the time average of any property A calculated along its trajectory is the same as the ensemble average of this property:

$$\langle A \rangle = \frac{1}{t} \int_0^t A(\mathbf{x}(t')) dt' = \int A(\mathbf{x}) \rho(\mathbf{x}) d\mathbf{x} \quad (12)$$

A system can only be ergodic if, starting from any given configuration, all other configurations can and will be reached in the course of the (sufficiently long) trajectory. Irreducibility ensures that any state can be reached from any other, but not necessarily that it will be reached; i.e., irreducibility is a necessary but not sufficient condition for ergodicity. Intuitively, the fact that a reducible transition matrix cannot lead to ergodic dynamics arises from the observation that a set of n elements in $\mathbf{p}(t)$ will never undergo transitions to the complementary set of $N_d - n$ elements in $\mathbf{p}(t + \tau)$. As a result, ensemble dynamics initiated from a specific distribution $\mathbf{p}(0)$ solely encompassing nonvanishing

elements of the former set will never generate probabilities in the latter set.

According to the Perron–Frobenius theorem,^{33,34} if a column-stochastic transition matrix $\mathbf{T}(\tau)$ is irreducible, it additionally has the following properties:

3. Its eigenvalue $\lambda_1 = 1$ is nondegenerate; i.e., it is associated with a unique real right eigenvector $\boldsymbol{\psi}_1$.
4. The components of the right eigenvector $\boldsymbol{\psi}_1$ are all non-negative.

The unique right eigenvector $\boldsymbol{\psi}_1$ associated with the eigenvalue $\lambda_1 = 1$ is referred to as the *stationary* probability distribution of the ensemble dynamics and will be further noted as $\boldsymbol{\pi}$. Its properties are

$$\mathbf{T}(\tau)\boldsymbol{\pi} = \boldsymbol{\pi} \quad \forall \tau \quad (13)$$

and

$$\begin{aligned} \pi_i &\in \mathbb{R} \quad \forall i \\ \pi_i &\geq 0 \quad \forall i \\ \sum_i \pi_i &= 1 \end{aligned} \quad (14)$$

Intuitively, $\boldsymbol{\pi}$ corresponds to a special probability distribution within the ensemble that is invariant upon propagation by $\mathbf{T}(\tau)$ for any lag time τ .

At this point, one may add a fourth assumption to the assumptions underlying irreducible Markovian dynamics, namely that of primitivity. A non-negative square matrix \mathbf{A} is called *primitive* if there exists an integer $k > 0$ for which all elements of the matrix \mathbf{A}^k are positive. A sufficient condition for a non-negative and irreducible square matrix to be primitive is that it possesses at least one nonzero element on the diagonal. If $\mathbf{T}(\tau)$ is irreducible and has at least one positive entry on its diagonal, then there is only one eigenvalue with $|\lambda_\alpha| = 1$ and this is $\lambda_1 = 1$; i.e., $\lambda_1 = 1$ is the only eigenvalue on the unit circle. The condition of primitivity ensures that in the limit of long lag times any arbitrary initial probability distribution $\mathbf{p}(0)$ converges to the stationary distribution $\boldsymbol{\pi}$.³⁴

$$\lim_{\tau \rightarrow \infty} \mathbf{T}(\tau) \mathbf{p}(0) = \boldsymbol{\pi} \quad \forall \mathbf{p}(0) \quad (15)$$

As a last stipulation, we require that Markovian dynamics (as defined by eq 1) are *detailed balanced* with respect to their stationary distribution. This is the case when

$$T_{ij}(\tau) \pi_j = T_{ji}(\tau) \pi_i \quad \forall i, j, \tau \quad (16)$$

or introducing the diagonal matrix $\boldsymbol{\Pi}$ with elements equal to $\boldsymbol{\pi}$, i.e., $\Pi_{ij} = \pi_i \delta_{ij}$:

$$\mathbf{T}(\tau) \boldsymbol{\Pi} = \boldsymbol{\Pi} \mathbf{T}^T(\tau) \quad \forall \tau \quad (17)$$

where \mathbf{T}^T denotes the transpose of the matrix \mathbf{T} . Detailed balance implies that the number of transitions between pairs of configurational points in a stationary ensemble (i.e., characterized by the probability distribution $\boldsymbol{\pi}$) is equal in the forward and backward directions. When the ensemble dynamics satisfy this condition, the stationary distribution $\boldsymbol{\pi}$ will be further referred to as the *equilibrium* probability distribution or Boltzmann distribution of the ensemble. Intuitively, a violation of detailed balance implies that for at least one pair of configurational points, there exists a net direct flow in the forward direction from the first point to the second that must be compensated by an equivalent net indirect flow via other points in the opposite direction to maintain the probability stationary. In the language of thermodynamics, this

behavior is characteristic of a steady state rather than an equilibrium stationary situation as would be encountered, e.g., in a system where a temperature, pressure, or composition gradient is maintained. At thermodynamic equilibrium, direct flows in the forward and backward directions between all pairs of states must compensate for each other, as will be the case, e.g., in a system where temperature, pressure, and composition are homogeneous in space.

Note that if $\mathbf{T}(\tau)$ is detailed-balanced, then so is any transition matrix $\mathbf{T}(n\tau) = \mathbf{T}^n(\tau)$ with $n \in \mathbb{Z}$. Note also that a column-stochastic matrix can only be detailed-balanced with respect to a vector that is also a right eigenvector associated with the eigenvalue one. In other words, irreducible Markovian dynamics can only be detailed balanced with respect to $\boldsymbol{\pi}$ (and no other vector). If an irreducible and primitive column-stochastic transition matrix $\mathbf{T}(\tau)$ is detailed balanced with respect to its stationary distribution $\boldsymbol{\pi}$, it also has the following properties:³⁴

- 1 All eigenvalues are real and lie in the interval $]-1; +1]$, so that all eigenvectors are real.
- 2 The eigenvectors of $\mathbf{T}(\tau)$ define a complete eigenbasis being orthonormal with respect to a weighted inner product.

The detailed balance condition has a number of very pleasant implications. First, the transition matrix becomes easier to grasp in terms of physical intuition because one is relieved from the necessity to find a physical interpretation for complex eigenvectors and eigenvalues. Second, since the eigenvectors of a detailed balanced and irreducible transition matrix $\mathbf{T}(\tau)$ form a complete basis of \mathbb{R}^{N_d} , where N_d is the dimension of the transition matrix, any vector $\mathbf{p}(t)$ can be expressed as a linear combination of these eigenvectors:

$$\mathbf{p}(t) = \sum_{\alpha} k_{\alpha}(t) \boldsymbol{\psi}_{\alpha} = \sum_{\alpha} c_{\alpha} \lambda_{\alpha}(t) \boldsymbol{\psi}_{\alpha} \quad (18)$$

After time $n\tau$, $\mathbf{p}(t + n\tau)$ is given as

$$\mathbf{T}(n\tau) \mathbf{p}(t) = \mathbf{T}^n(\tau) \mathbf{p}(t) = \mathbf{p}(t + n\tau) \quad (19)$$

Using

$$\mathbf{p}(t + n\tau) = \sum_{\alpha} c_{\alpha} \lambda_{\alpha}^n(t) \boldsymbol{\psi}_{\alpha} \quad (20)$$

the probability distribution $\mathbf{p}(t)$ can be interpreted as consisting of modes $\{\boldsymbol{\psi}_{\alpha}\}$ which show a temporal behavior according to the corresponding eigenvalues $\{\lambda_{\alpha}\}$. More precisely, the temporal behavior is an exponential decay, as can be seen from the following transformation:

$$\begin{aligned} \lambda_{\alpha}(t = n\tau) &= \lambda_{\alpha}^n(\tau) = \lambda_{\alpha}^{t/\tau}(\tau) = \exp(\ln(\lambda_{\alpha}^{t/\tau}(\tau))) \\ &= \exp\left(\frac{t}{\tau} \ln(\lambda_{\alpha}(\tau))\right) = \exp\left(-\frac{t}{\mu_{\alpha}}\right) \end{aligned} \quad (21)$$

where $t = n\tau$ and the mean lifetime μ_{α} is given as

$$\mu_{\alpha} = -\frac{\tau}{\ln(\lambda_{\alpha}(\tau))} \quad (22)$$

In the context of Markov models, μ_{α} is typically referred to as the *implied time scale* of the decay process. The mode that corresponds to the eigenvalue $\lambda_1 = 1$ does not decay, which can be seen either by realizing that $\lambda_1^n = \lambda_1 = 1$ for any n or by noting that the argument in the exponential in eq 21 becomes 0 if $\lambda_{\alpha} = 1$. This result corresponds to the earlier result that the eigenvector associated with λ_1 is the equilibrium distribution: $\boldsymbol{\psi}_1 = \boldsymbol{\pi}$. The

other modes, which all correspond to an eigenvalue with $|\lambda_\alpha| < 1$, will vanish for $n \rightarrow \infty$. The smaller the value of λ_α , the faster the corresponding mode will decay.

The derivation of the implied time scales has been based on the assumption that $\mathbf{T}(\tau)$ represents a Markov process. In this case, the eigenvector expansion can be based on $\mathbf{T}(\tau)$ or any other $\mathbf{T}(n\tau)$ and will lead to the same implied time scale, i.e.,

$$\mu_\alpha = \frac{-n\tau}{|\lambda_\alpha(n\tau)|} = \text{const } n = 1, 2, \dots \quad (23)$$

Conversely, eq 23 can be used as a check for Markovian behavior. Plotting the implied time scales of transition matrices with various lag times $n\tau$ yields a set of constant functions if the underlying dynamics are Markovian.^{7,8,11}

In the limit $\tau \rightarrow \infty$, all eigenmodes except for the stationary one have decayed, and there must be a matrix

$$\mathbf{T}_1 \doteq \lim_{\tau \rightarrow \infty} \mathbf{T}(\tau) \quad (24)$$

which immediately returns the equilibrium distribution $\boldsymbol{\psi}_1 = \boldsymbol{\pi}$ when multiplied by any arbitrary distribution $\mathbf{p}(t)$, i.e.

$$\mathbf{T}_1 \mathbf{p}(t) = \boldsymbol{\pi} \quad (25)$$

We will call the matrix \mathbf{T}_1 the *equilibrium matrix*.

3. MODELS

3.1. Bit-Flip Model. The coupling between the environment and a system can be studied on a simple bit-flip model consisting of two bits, S and E . Bit S represents the system and bit E the environment. Either of these bits can assume two states: \uparrow ("up") or \downarrow ("down"). The time scale of the dynamics of these two bits is not the same and is determined by their flipping probabilities p_S and p_E , respectively. The two bits can be coupled or uncoupled.

The complete system (consisting of S and E) has four states: $\uparrow\uparrow$ (state 1), $\uparrow\downarrow$ (state 2), $\downarrow\uparrow$ (state 3), and $\downarrow\downarrow$ (state 4), where the first arrow stands for bit S and the second for bit E . The dynamics of the complete system are modeled using a 4×4 transition matrix $\mathbf{T}_{\text{bit}}(\tau)$ which contains the transition probabilities between those four states. To obtain a transition matrix which only represents the dynamics of S , we project the matrix \mathbf{T}_{bit} onto the states of bit S . This procedure intrinsically assumes that the dynamics of S are Markovian, where we identified S with \mathbf{x} and E with \mathbf{y} . Using the implied time scales as a measure for Markovian behavior, we can study how a coupling between E and S violates this assumption.

The transition matrix of the complete system dynamics $\mathbf{T}_{\text{bit}}(\tau)$ is constructed in the following fashion. The probability that the bit S will flip, i.e., that it will make a transition $\uparrow \rightarrow \downarrow$ or a transition $\downarrow \rightarrow \uparrow$, within time τ is given by p_S , and the probability that it will stay in its current state is given as $1 - p_S$. Consequently, the transition matrix for a single bit S (bit E not present) is given as

$$\mathbf{T}_S(\tau) = \begin{pmatrix} 1 - p_S & p_S \\ p_S & 1 - p_S \end{pmatrix} \quad (26)$$

where $\mathbf{T}_{S,11}$ represents the transition probability for $\uparrow \rightarrow \uparrow$, $\mathbf{T}_{S,21}$ the transition probability for $\uparrow \rightarrow \downarrow$, $\mathbf{T}_{S,12}$ the transition probability for $\downarrow \rightarrow \uparrow$, and $\mathbf{T}_{S,22}$ the transition probability for $\downarrow \rightarrow \downarrow$. Likewise the transition matrix for a single bit E (bit S not present) is given as

$$\mathbf{T}_E(\tau) = \begin{pmatrix} 1 - p_E & p_E \\ p_E & 1 - p_E \end{pmatrix} \quad (27)$$

where p_E denotes the probability that bit E will flip within τ . The transition for the complete system consisting of the two non-interacting bits is given by the Kronecker product (sometimes called tensor product) of \mathbf{T}_S and \mathbf{T}_E :

$$\begin{aligned} \mathbf{T}_{\text{bit}}(\tau) &= \mathbf{T}_S(\tau) \otimes \mathbf{T}_E(\tau) \\ &= \begin{pmatrix} (1-p_S)(1-p_E) & (1-p_S)p_E & p_S(1-p_E) & p_S p_E \\ (1-p_S)p_E & (1-p_S)(1-p_E) & p_S p_E & p_S(1-p_E) \\ p_S(1-p_E) & p_S p_E & (1-p_S)(1-p_E) & (1-p_S)p_E \\ p_S p_E & p_S(1-p_E) & (1-p_S)p_E & (1-p_S)(1-p_E) \end{pmatrix} \end{aligned} \quad (28)$$

This matrix represents transitions between states $\uparrow\uparrow$ (state 1), $\uparrow\downarrow$ (state 2), $\downarrow\uparrow$ (state 3), and $\downarrow\downarrow$ (state 4) where the first arrow stands for bit S and the second for bit E . A coupling is introduced into the dynamics of the complete system by selectively modifying elements of $\mathbf{T}_{\text{bit}}(\tau)$ and renormalizing its columns.

In order to obtain the transition matrix of bit S , one needs to project $\mathbf{T}_{\text{bit}}(\tau)$ onto the state of bit S using a projection

$$\mathbf{T}_{S,\text{proj}}(\tau) = \mathbf{P}^T \mathbf{T}_{\text{bit}}(\tau) \mathbf{P} \quad (29)$$

with

$$\mathbf{P} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \quad (30)$$

Note that if \mathbf{T}_{bit} represents the dynamics of the two uncoupled bits, i.e., if $\mathbf{T}_{\text{bit}}(\tau) = \mathbf{T}_S(\tau) \otimes \mathbf{T}_E(\tau)$, then eq 29 recovers $\mathbf{T}_S(\tau)$.

In the present application of the bit-flip model, a coupling between the environment E and the system S was introduced by multiplying the elements T_{11} , T_{14} , T_{41} , and T_{44} of \mathbf{T}_{bit} by a (positive) coupling factor k and renormalizing the columns. This increases the probability of states in which the two spins are aligned ($\uparrow\uparrow$ and $\downarrow\downarrow$). The coupling constant k was varied between 1 (no coupling) and 100 (strong coupling). The flipping probability of system S was set to $p_S = 0.100$. In order to examine the influence of the relaxation time of the environment on the system, the flipping probability of E , p_E , was varied between 0.001 (very slow dynamics, long relaxation time) and 0.150 (dynamics of the environment faster than the dynamics of the system, short relaxation time).

3.2. Butane Model. The other test system is liquid butane. We performed molecular dynamics (MD) simulations of boxes of 512 butane molecules at various temperatures ($T = 298.15$ K and $T = 400.00$ K) and densities (d : 50–500 u/nm³, where u denotes the atomic mass unit). For each density, a box with regularly placed butane molecules was constructed using the program `build_box` of GROMOS++,³⁵ which was then heated to the target temperature over a period of 10⁵ time steps (200 ps). For each system, a trajectory of 2 ns was generated using the GROMOS05 software,³⁵ which implements the leapfrog integrator,³⁶ and the GROMOS 45A3 force field.³⁷ All bond lengths were constrained using the SHAKE algorithm³⁸ with a relative tolerance of 10⁻⁴, allowing for a time step of 2 fs. Configurations of all 512 molecules were saved every 0.08 ps. The system was simulated in a rectangular box using periodic boundary conditions. The volume was kept constant, and the molecules were weakly coupled to one temperature bath of 298.15 K or 400.00 K³⁹ with a coupling time of 0.1 ps. We used

Table 1. Overview of the Simulations Performed

T (K)	MD setup				SD setup			
	density, ρ (u/nm ³)	number of molecules	simulation length (ns)	D (σ) (10^{-2} nm ² /ps)	γ_{friction} (1/ps)	number of molecules	simulation length (μ s)	
298.15	300	512	2	1.403(1.224)	5.0	1	1	
298.15	345	512	2	0.7113(0.5612)	10.0	1	1	
298.15	400	512	2	0.3952(0.2815)	17.9	1	1	
298.15	450	512	2	0.1981(0.2009)	35.8	1	1	
298.15	500	512	2	0.0561(0.0469)	126.3	1	1	
400.00	50	512	2	18.18(16.31)	0.5	1	1	
400.00	100	512	2	8.904(7.293)	1.1	1	1	
400.00	150	512	2	5.273(5.009)	1.8	1	1	
400.00	200	512	2	4.075(3.610)	2.3	1	1	
400.00	250	512	2	2.630(2.502)	3.6	1	1	
400.00	300	512	2	1.501(1.446)	6.3	1	1	
400.00	345	512	2	1.321(1.134)	7.2	1	1	

0.8 nm/1.4 nm as a twin-range cutoff and 1.4 nm as a reaction field cutoff with $\epsilon_{\text{rf}} = 1.0$. The atom pair list for short-range interactions and the intermediate-range forces were updated every five steps.

We performed stochastic dynamics (SD) simulations of a single butane molecule at two different temperatures, 298.15 K and 400.00 K, using various friction coefficients, γ_{fric} : 0.5–126.3 ps⁻¹. The friction coefficients were calculated from the diffusion constants obtained from the above-described MD simulations using the relation:

$$\gamma_{\text{fric}} = \frac{k_{\text{B}}T}{Dm_{\text{solv}}} \quad (31)$$

where k_{B} is the Boltzmann constant, T is the temperature of the MD simulation, D is the diffusion constant, and m_{solv} is the mass of butane (58.124 g/mol). The diffusion constant was calculated as an average of the diffusion constants of 50 arbitrarily picked butane molecules where each diffusion constant was estimated from a least-squares fit to the Einstein equation:

$$D = \lim_{t \rightarrow \infty} \frac{\langle [\mathbf{r}_0 - \mathbf{r}(t)]^2 \rangle}{2Nt} \quad (32)$$

Here, \mathbf{r}_0 is the center of geometry of the first configuration in the trajectory, $\mathbf{r}(t)$ is the center of geometry at time t , and N is the number of dimensions taken into account, which was set to 3.

As in the MD simulations, all bonds were constrained using the SHAKE algorithm,³⁸ and a time step of 2 fs was used. Each system was simulated for 1 μ s, and the configuration of the molecule was saved every 0.1 ps. Vacuum boundary conditions were applied, and the temperature was maintained by the stochastic dynamic integrator.

A summary of all performed simulations, the obtained diffusion constants, and the corresponding friction coefficients is reported in Table 1.

3.3. Generation of the Transition Matrices $\mathbf{T}(\tau)$ for the Test System Butane. We consider a single butane immersed in a solvent of butane molecules where the solvent is modeled either explicitly using MD or implicitly using SD. The dominant degree of freedom of butane is the C₁–C₂–C₃–C₄-dihedral angle, which we discretize into equally sized microstates (bins). For most analyses, we use a bin size of 5° (72 bins per dihedral angle); only in Figure 6, we varied the bin size from 5° to 120° (72 to 3

bins per dihedral angle). For various values of the lag time τ (ranging from 80 fs to 100 ps), the configurations at time $t = 0, \tau, 2\tau, 3\tau$, etc. are mapped onto the microstates, and the transitions from microstate i to microstate j for each combination of i and j are counted. We enforce detailed balance by counting each transition $i \rightarrow j$ also as a transition $j \rightarrow i$. This “backward counting” inherently assumes that the trajectory is in global equilibrium and the deviation from detailed balance is only due to statistical errors. The numbers of transitions are sorted into a matrix, and the columns of the matrix are normalized by the total number of transitions in each column to obtain the column-stochastic transition matrix. When constructing the transition matrix from a MD trajectory, we regard one butane molecule as solute and the remaining 511 as solvent and count the transitions of the solute. Since the choice of the solvent molecule is arbitrary, we repeat this procedure 511 times, where in each round another molecule represents the solute. The transition matrix for a single butane is then constructed from the added transition counts of all 512 evaluations of the MD trajectory.

The implied time scales μ_i of each transition matrix $\mathbf{T}(\tau)$ are calculated as

$$\mu_i(\tau) = -\frac{\tau}{\ln |\lambda_i(\tau)|} \quad (33)$$

where τ is the time step of the transition matrix and $|\lambda_i|$ is the absolute value of its i th eigenvalue. We plot the implied time scales of the dominant eigenvalues and evaluate the reference implied time scales and τ_{Markov} by visual inspection.

4. RESULTS

We use two model systems—(i) a single butane immersed in a solvent of butane and (ii) a bit-flip model as described in section 3—to illustrate some important properties of transition matrices and to study the effect of marginal degrees of freedom on the dynamics of the relevant degrees of freedom.

4.1. Colormaps of Transition Matrices. Figures 1 and 2 show colormaps of various transition matrices of the dihedral angle degree of freedom of butane. The dihedral angle has been discretized into 72 microstates of 5° per microstate, and each point in the colormaps represents a transition probability T_{ij} from microstate j to microstate i . A high transition probability is marked in red, and a transition probability close to zero is marked

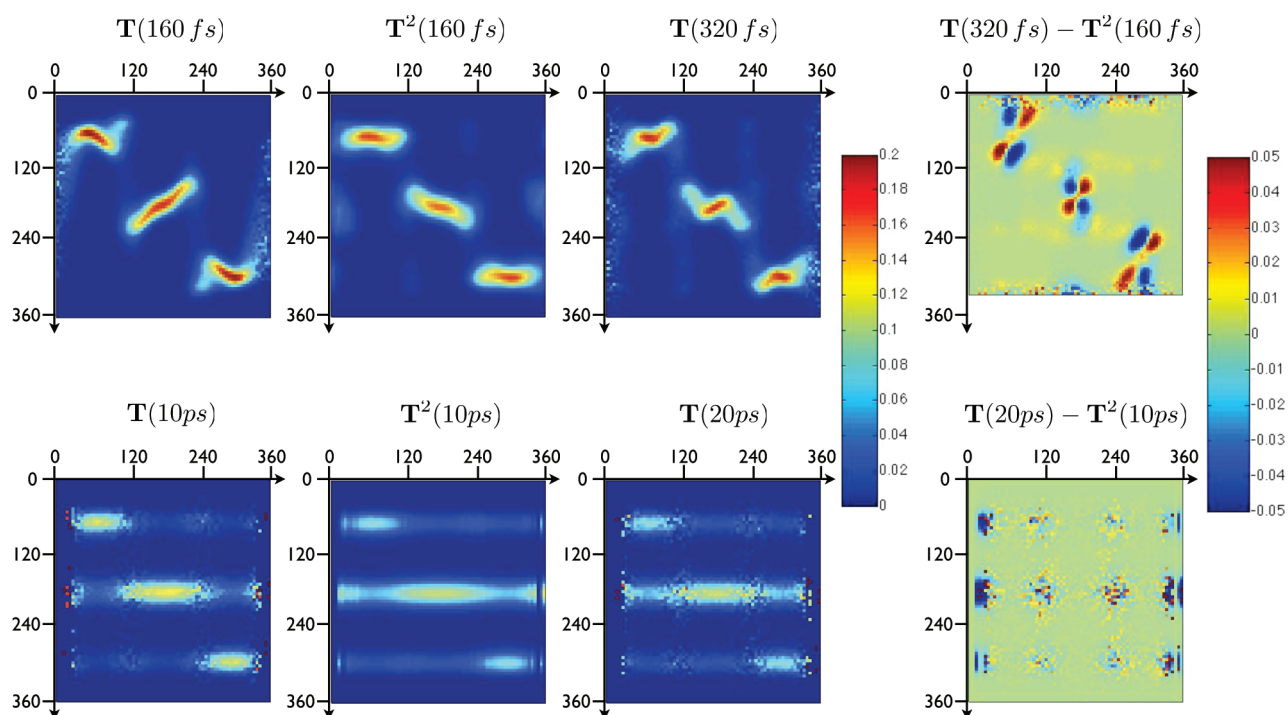


Figure 1. Transition matrices for the dihedral angle of a single butane immersed in butane ($T = 298.15$ K, $\rho = 345$ u/nm³). Upper row: non-Markovian regime, $\tau = 160$ fs and $\tau = 320$ fs. Lower row: Markovian regime, $\tau = 10$ ps and $\tau = 20$ ps. Right-most column: difference plots of $T(2\tau) - T^2(\tau)$.

in blue. The right-most column in Figure 1 and the lower row in Figure 2 show colormaps of difference matrices in which negative elements are marked blue, elements close to zero are green, and positive elements are red. Taking the first matrix in the upper row of Figure 1 as an example, one can clearly distinguish the three metastable states of butane as three red areas along the diagonal of the matrix. For any initial microstate in the *gauche* state of butane between 0° and 120° , the molecule has a high transition probability to any microstate within this state and a low transition probability to any microstate outside this state within lag time τ . The microstates between 0° and 120° are said to be “kinetically close”. Analogously, microstates which correspond to the *trans* state (120° – 240°) are kinetically close, as are microstates which correspond to the second *gauche* state (120° – 360°).

4.2. Illustration of the Chapman–Kolmogorov Equation.

Equation 9, according to which taking a transition matrix with lag time τ to the power n yields a matrix which is equal to a transition matrix with n times longer lag time if the dynamics are Markovian, is illustrated in Figure 1 using transition matrices of butane as an example. The first three columns show colormaps of transition matrices, and the fourth column shows colormaps of difference matrices.

Transition matrices with short lag times on the order of a few hundred femtoseconds are depicted in the upper row. The second matrix is the square of the first one with lag time $\tau = 160$ fs and should be equal to the third matrix, which has been constructed with a lag time of $\tau = 2 \times 160$ fs = 320 fs if the dynamics are Markovian at this time resolution of the model. This is clearly not the case, as the second and the third matrix already visually differ from each other. The difference matrix correspondingly shows systematic deviations from zero. If one was to evolve a density with $T^2(160$ fs), its dynamics would systematically deviate from the dynamics of the same density evolved with $T(320$ fs).

The lower row shows transition matrices with longer lag times on the order of 10 ps. In this time regime, the dynamics of the dihedral angle can be approximated by a Markov process. $T^2(10$ ps) and $T(20$ ps) are visually similar, except for the fact that $T(20$ ps) shows more noise than $T^2(10$ ps). This is due to the poorer sampling for longer lag times. Accordingly, the difference matrix depicted in the right-most column shows no systematic deviations from zero but only deviations which are due to the noise in the two matrices. Note that the amplitude of noise varies with the magnitude of the transition probabilities in $T^2(10$ ps) and $T(20$ ps).

4.3. Illustration of the Equilibrium Matrix. Figure 2 illustrates the concept of the equilibrium matrix T_1 defined in eq 25 using transition matrices of a butane molecule with a lag time of 5 ps as an example. For this lag time, the dynamics of the system are Markovian and the equilibrium matrix T_1 can be constructed from the first eigenvector of the transition matrix, which is equal to the equilibrium distribution. It is depicted in the right-most panel, and as its columns are all equal to the equilibrium distribution, its colormap shows a striped pattern. When multiplied by an arbitrary initial distribution, it returns the equilibrium distribution. T_1 does not contain any information on the kinetic proximity of groups of microstates, and metastable states cannot be extracted from this matrix.

The other two panels in the upper row of Figure 2 show the transition matrix with lag time 5 ps, $T(5$ ps), and the square of it, $T^2(5$ ps), which is approximately equal to $T(10$ ps). At lag time $\tau = 5$ ps (left-most panel), the three metastable states of butane are clearly discernible, implying that the equilibration time of the system is longer than 5 ps. In the middle panel, the metastable states are less discernible, and the stripe pattern emerges. At time $t = \tau = 10$ ps, the probability of finding the system in any of the three metastable states is still slightly biased toward finding it in its initial state. But the information about the kinetic proximity of the microstates contained in the matrix is less clear. The lower

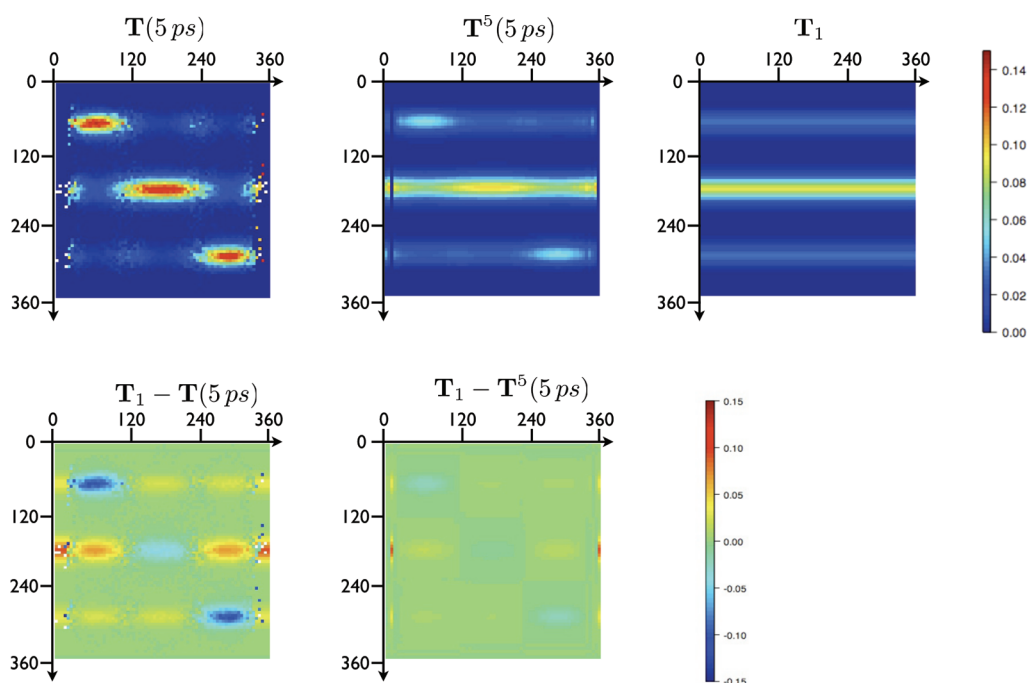


Figure 2. Transition matrices for the dihedral angle of a single butane immersed in butane ($T = 298.15$ K, $\rho = 345$ u/nm³). Upper row: $T(5$ ps), $T^5(5$ ps), and equilibrium matrix T_1 constructed from the first eigenvector ψ_1 of $T(5$ ps). Lower row: difference matrices between the equilibrium matrix and $T(5$ ps) and $T^5(5$ ps), respectively.

row shows the difference matrices $T_1 - T(5$ ps) and $T_1 - T^2(5$ ps). For $\tau = 5$ ps (left most panel), the transition matrix shows large systematic deviations from the equilibrium matrix. For $\tau = 10$ ps (middle panel), we see the same deviations which are, however, much smaller in absolute value.

4.4. Coupling of Marginal and Relevant Degrees of Freedom. When constructing Markov models from MD simulation data, the complete phase space is split into relevant degrees of freedom for which the model is constructed and marginal degrees of freedom which are assumed to act as stochastic forces on the relevant degrees of freedom. Depending on the time scale of the dynamics of the marginal degrees of freedom and the strength of the coupling between the marginal and the relevant degrees of freedom, this assumption can be fulfilled to a greater or lesser extent. In the bit-flip model, the relevant degrees of freedom are modeled by the bit S and the marginal degrees of freedom by the bit E . The time scale of the dynamics of E is determined by the flipping probability p_E : the higher the p_E , the faster the dynamics. The strength of the coupling is determined by the coupling constant k . In all applications of the bit-flip model presented here, the flipping probability of the system, p_S , was set to 0.100.

In Figure 3, the influence of the flipping probability on the implied time scale of the second eigenvalue of $T_{S,proj}$ is illustrated. The brown curve ($p_E = 0.150$) represents the case in which the dynamics of the environment is faster than the dynamics of the system. The implied time scale of the projected matrix rises until it reaches a plateau at about $n = 150$. This is the threshold in time resolution τ_{Markov} after which the dynamics of the system are Markovian until, after $n = 750$, the curve diverges again from a constant implied time scale. The latter deviation is caused by the fact that for a very high number of iterations the transition matrix approaches the equilibrium matrix, and the second eigenvector becomes so small that it is susceptible to numerical errors. Note that the system S has a flipping probability of 1–10 within a time

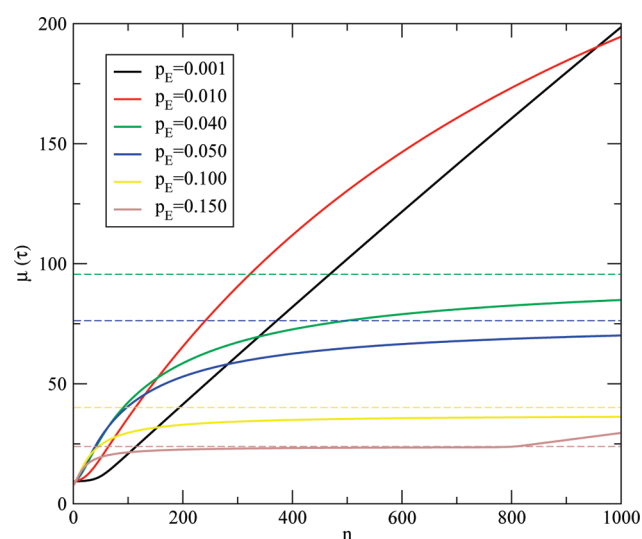


Figure 3. Implied time scale μ of the bit-flip model as a function of lag time $n\tau$ calculated from the second eigenvalue of $T_{S,proj}(n\tau)$ as a function of n where T_{bit} has been constructed with varying the flipping probability, p_E , of E . Coupling constant $k = 100$. Flipping probability of the system $p_S = 0.100$. Thin dashed lines: true implied time scales, μ_2 , as calculated from the respective T_{bit} . For $p_E = 0.001$, $\mu_2 = 3828.7$, and for $p_E = 0.010$, $\mu_2 = 383.5$, which is well beyond the region shown.

step τ . If the dynamics of the system are modeled with a time resolution of 150τ , a single iteration does not yield the probability of a *single* transition between the two states but the probability of finding the system in one of the states after a *sequence* of transitions. Each transition in this sequence has occurred under the influence of the environment; i.e., it was not Markovian, but the long lag time of the projected model provides for an ensemble of transitions in which the influence of the environment averages

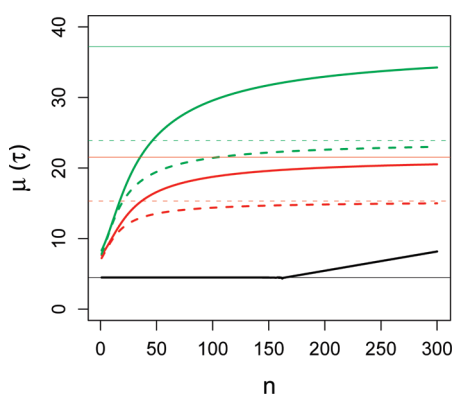


Figure 4. Implied time scale μ of the bit-flip model as a function of the lag time $n\tau$ calculated from the second eigenvalue of $\mathbf{T}_{S,\text{proj}}(n\tau)$ as a function of n where \mathbf{T}_{bit} has been constructed with varying values of the coupling constant k and the flipping probability of the environment p_E . Flipping probability of the system $p_S = 0.100$. Solid black: $k = 1$, $p_E = 0.100$. Dashed black: $k = 1$, $p_E = 0.150$. Solid red: $k = 10$, $p_E = 0.100$. Dashed red: $k = 10$, $p_E = 0.150$. Solid green: $k = 100$, $p_E = 0.100$. Dashed green: $k = 100$, $p_E = 0.150$. The solid black and dashed black lines are on top of each other. Thin lines: corresponding true implied time scales μ_2 as calculated from the respective \mathbf{T}_{bit} .

out. The faster the dynamics of the environment, the smaller the sequence of transitions has to be until the influence of the environment is averaged out, i.e., the smaller τ_{Markov} will be. Unless the time scale of the environment and the time scale of the system differ by orders of magnitude, a Markov model of the system dynamics does not represent the probability of a single transition within lag time τ_{Markov} but the probability of finding the system in state j at time $t = \tau_{\text{Markov}}$ given that it started its sequence of transitions in state i at time $t = 0$.

The yellow curve in Figure 3 represents the case in which the dynamics of the environment have the same time scale as the dynamics of the system. The implied time scale becomes approximately constant after 250 iterations, at which time the transition matrix is, however, so close to the equilibrium matrix that the model does not contain any significant information on the dynamics of the system. Similarly, the green ($p_E = 0.040$) and the blue ($p_E = 0.050$) curves represent cases in which the dynamics of the environment have a time scale which is on the same order of magnitude as the time scale of the system but slightly larger. In this case, the implied time scale curves slowly level off but never reach a plateau region. If one encounters this type of behavior when constructing a Markov model from MD simulation data, one should consider to include more degrees of freedom into the Markov model.

Finally, the black curve ($p_E = 0.001$) and the red curve ($p_E = 0.010$) in Figure 3 represent the case in which the dynamics of the environment are much slower than the dynamics of the system. For few iterations of the complete transition matrix, $n < 30$ and $n < 10$, respectively, the environment has hardly changed, and therefore, the system does not feel its influence. The implied time scales are constant at about a value of 10. After about 10–30 iterations, the environment has changed noticeably from its state at $t = 0$ and starts to influence the dynamics of the system. However, because the dynamics of the environment are so slow, even 1000 iterations are not sufficient to provide enough statistics to average out the influence of the environment on the system, and the curve never reaches a plateau region (data not shown).

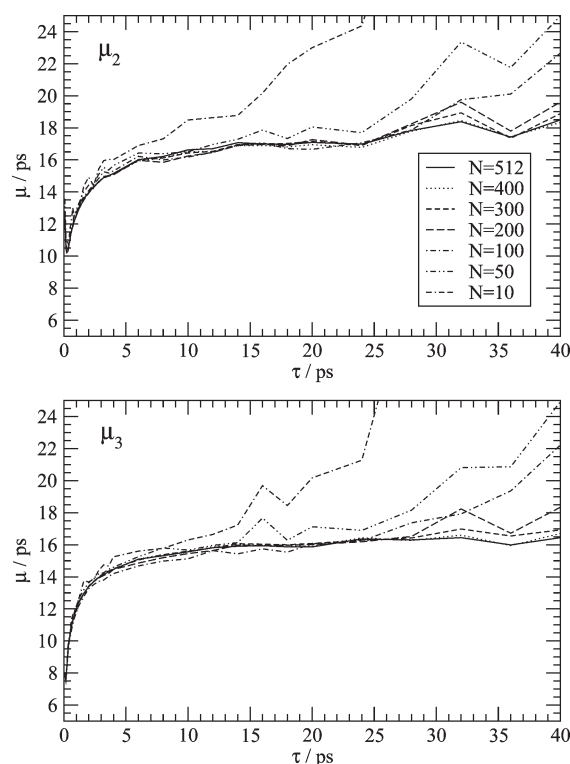


Figure 5. Implied time scale as a function of lag time τ for various numbers of data points used in the analysis. Implied time scales μ_2 and μ_3 calculated from the MD simulation of one butane molecule immersed in liquid butane at $T = 298.15$ K and $\rho = 345$ u/nm³. The number of molecules N used for the analysis of a total of 512 molecules varied from 10 to 512.

Figure 4 shows the influence of the coupling constant on the dynamics of the system for two time scales of the environment $p_E = 0.100$ and $p_E = 0.150$. If there is no coupling, i.e., if $k = 1$, the dynamics of the system are independent of the environment, and consequently, the implied time scales are constant (black curves). As before, the deviation from constant μ after $n = 150$ is due to numerical errors. Since the transition probabilities are independent of the state of the environment, the implied time scales are also independent from the flipping probability of E . Both have a value of 4.48. Raising the coupling constants to $k = 10$ and $k = 100$ changes the implied time scale of the system, in this case, raising it. We note, however, that this might be caused by the choice of matrix elements which are modified by k . τ_{Markov} is larger the stronger the coupling between the environment and the system is. For a flipping probability of $p_E = 0.100$, $\tau_{\text{Markov}} \approx 100\tau$ for $k = 10$ and $\tau_{\text{Markov}} \approx 150\tau$ for $k = 100$. For a flipping probability of $p_E = 0.150$, $\tau_{\text{Markov}} \approx 170\tau$ for $k = 10$ and $\tau_{\text{Markov}} > 300\tau$ for $k = 100$.

4.5. Behavior of Liquid Butane. As a test system, we consider a single butane immersed in a solvent of butane at two temperatures, $T = 298.15$ K and $T = 400.00$ K, and various densities. In one set of simulations, we model the solvent explicitly with 511 butane molecules; in the other set of simulations, we model the solvent implicitly using stochastic dynamics. The dominant degree of freedom in butane is the dihedral angle between its carbon atoms $C_1-C_2-C_3-C_4$, which we use for the construction of the Markov model. All other degrees of freedom (bond-angle vibrations and solvent degrees of freedom in MD) are marginal in the model and are assumed to interact stochastically

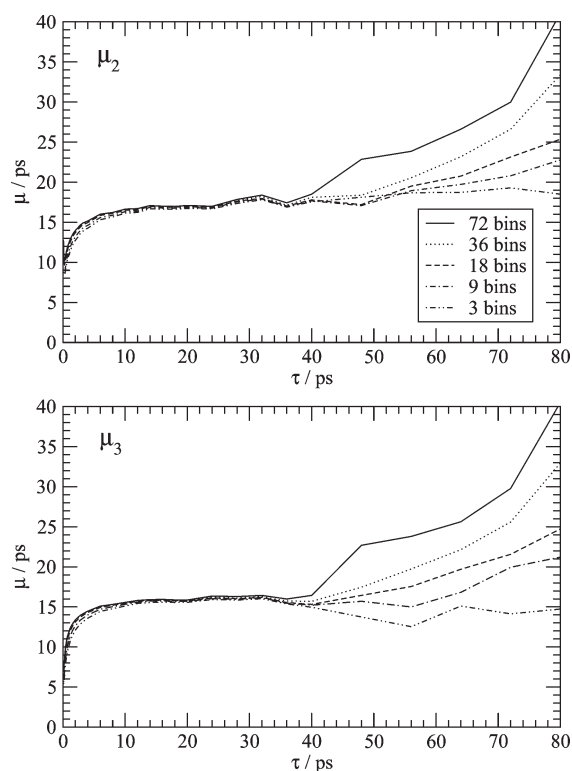


Figure 6. Implied time scale as a function of lag time τ for various resolutions of the configuration space. Implied time scales μ_2 and μ_3 calculated from the MD simulation of 1 butane molecule immersed in liquid butane at $T = 298.15$ K and $\rho = 345$ u/nm³. The number of microstates (bins per dihedral angle) varied from 3 to 72.

with the dihedral-angle degree of freedom. The molecule has three metastable states, represented by the *gauche+*, *trans*, and *gauche-* conformation of the dihedral angle. Correspondingly, it has two dominant eigenvalues, λ_2 and λ_3 ($\lambda_1 = 1$), which we used to calculate implied time scales and to determine τ_{Markov} .

When constructing a Markov model from simulation data, the upper bound of a possible lag time is not set by the numerical accuracy with which the eigenvalue can be calculated for a transition matrix approaching the equilibrium matrix but by the extent of the sampling. Because data points are evaluated at $t = 0, \tau, 2\tau$, etc., the longer the τ , the fewer data points are available in a trajectory of a given length. Figure 5 illustrates this fact. The first panel shows the implied time scale of the second eigenvalue, and the second panel shows it for the third eigenvalue calculated from MD simulations at $T = 298.15$ K and a density of $\rho = 345$ u/nm³. We have varied the number of data points, N , used for the construction of the Markov model by varying the number of times the MD trajectory is evaluated, where at each evaluation a different butane molecule was considered to be the solute. τ_{Markov} is not influenced by the amount of data the Markov model is built upon. It lies between $\tau = 5$ ps and $\tau = 10$ ps. However, the length of the plateau region is sensitive to the amount of data. The less data used, the smaller the lag time for which the implied time scales diverge from the plateau. In particular, if the trajectory is evaluated only 10 times, the implied time scales diverge before τ_{Markov} is reached.

Figure 6 illustrates how the resolution of the relevant degrees of freedom influences the implied time scale. For small lag times up to 40 ps, there is only a very small but systematic influence of

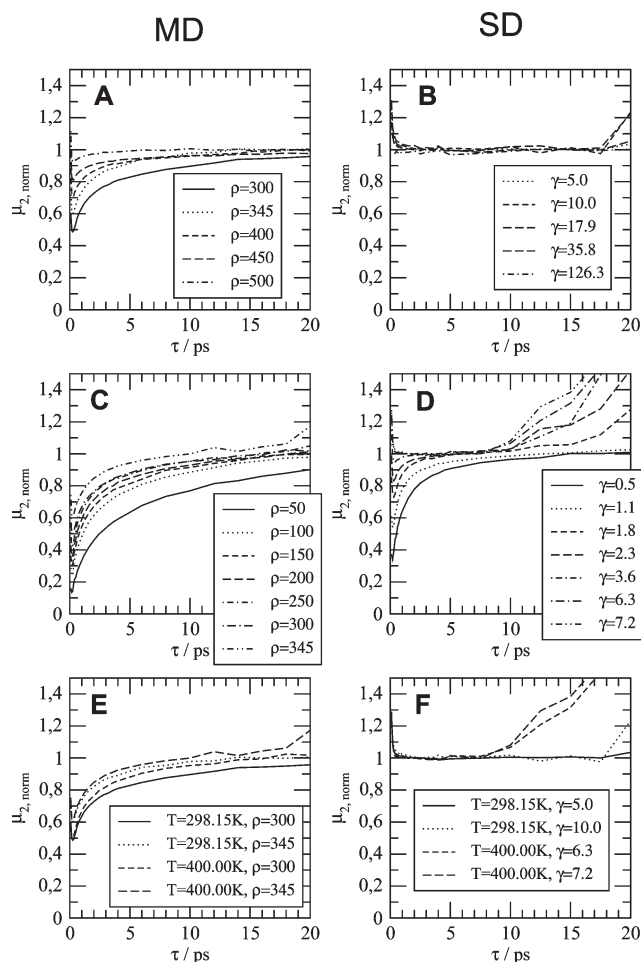


Figure 7. Normalized implied time scale $\mu_{2,\text{norm}}$ (eq 34) as function of the lag time τ for various systems. Left column: MD simulations of 1 butane molecule immersed in liquid butane. Right column: SD simulations of 1 butane molecule immersed in liquid butane modeled by the friction coefficient γ . Panel A: $T = 298.15$ K, density ρ varied from 300 u/nm³ to 500 u/nm³. Panel B: $T = 298.15$ K, γ varied from 5.0 ps⁻¹ to 126.3 ps⁻¹. Panel C: $T = 400.00$ K, ρ varied from 50 u/nm³ to 345 u/nm³. Panel D: $T = 400.00$ K, γ varied from 0.5 ps⁻¹ to 7.2 ps⁻¹. Panel E: Temperature and density varied. Panel F: Temperature and density varied.

the resolution of the discretization on the implied time scale curves that is discernible. More precisely, τ_{Markov} is only slightly smaller for finer discretizations. The figure also illustrates the effect of insufficient sampling of the transition probabilities. For lag times of $\tau = 40$ ps and greater, the number of available data points becomes so small that the statistical error on the transition probabilities is too large to yield a reliable Markov model. Consequently, the curves diverge from the plateau. The effect is the greater, the finer the resolution of the model. In practice, one can improve the sampling by using a sliding window, i.e., by counting the transition from *every* time step in the simulation to the time step τ further, instead of using only the time steps at 0, τ , 2τ , 3τ , etc., as done here.

We note that our discretization is special in two aspects. First, each metastable state corresponds to a single minimum in the free-energy surface and does not consist of several substates. If by lowering the resolution of the discretization, several states are grouped into one microstate, the lag time at which the model becomes Markovian does change.¹¹ Second, we ensured that

Table 2. Overview of the Implied Time Scales of the Second and Third Eigenvalues Observed in the Markovian Regime of the Dynamics for Various Temperatures and Densities^a

system			MD		SD	
<i>T</i> (K)	density (u/nm ³)	γ_{friction} (ps ⁻¹)	$\mu_{2,\text{reference}}$ (ps)	$\mu_{3,\text{reference}}$ (ps)	$\mu_{2,\text{reference}}$ (ps)	$\mu_{3,\text{reference}}$ (ps)
298.15	300	5.0	21.0	20.2	10.2	8.0
298.15	345	10.0	17.0	16.2	11.2	7.2
298.15	400	17.9	15.0	12.4	11.8	7.2
298.15	450	35.8	13.4	9.8	17.4	10.0
298.15	500	126.3	12.8	8.8	43.4	26.0
400.00	50	0.5	30.0	28.0	12.6	10.6
400.00	100	1.1	16.5	16.0	7.6	6.3
400.00	150	1.8	13.0	12.6	5.4	5.2
400.00	200	2.3	11.0	10.8	4.6	4.6
400.00	250	3.6	9.2	9.0	4.0	3.6
400.00	300	6.3	7.8	7.2	3.8	2.8
400.00	345	7.2	6.4	5.8	3.8	2.8

^a Column MD: explicit solvent model. Column SD: implicit solvent model.

there is a microstate boundary exactly at the peak of the free-energy barrier between the metastable states. Moving this boundary away from the barrier peak will decrease the quality of the Markov model.^{9,26} The error introduced by discretizing the relevant degrees of freedom has, however, a finite upper bound.⁹

In the butane test system, the marginal degrees of freedom are predominantly those of the solvent molecules, exceptions being the bond-angle degrees of freedom. Their coupling to the relevant degree of freedom, the dihedral angle, is determined by the model of the solvent, implicit or explicit; the density; and the temperature. Their influence on τ_{Markov} is examined in Figure 7.

Since model, density, and temperature do not only influence τ_{Markov} but also the implied time scales used to determine τ_{Markov} we introduce a normalized implied time scale:

$$\mu_{i,\text{norm}}(\tau) = \frac{\mu_i(\tau)}{\mu_{i,\text{reference}}} \quad (34)$$

where $\mu_i(\tau)$ indicates the implied time scale of the *i*th eigenvalue and $\mu_{i,\text{reference}}$ indicates the reference implied time scale, i.e., the time scale in the Markovian regime, which we determine by visual inspection. Table 2 lists the observed reference implied time scales. The column “MD” corresponds to an explicit solvent model; the column “SD” to an implicit one. The fact that stochastic dynamics (SD) underestimate the relaxation times of a system, i.e., underestimates the implied time scales, if the fundamental assumption underlying this type of dynamics, a large heavy particle in a solvent of small light particles, is not fulfilled, is a known effect. The expectation that the system equilibrates quicker if the temperature or the density is increased is reflected in the corresponding decrease of the implied time scale. The only exceptions to this trend are the simulations with very high friction coefficients ($\gamma_{\text{friction}} = 35.8 \text{ ps}^{-1}$ and $\gamma_{\text{friction}} = 126.3 \text{ ps}^{-1}$). In these cases, the velocity of the dihedral-angle degree of freedom is decreased so drastically at each simulation step that transitions between the metastable states are very rare. Consequently, the equilibration between these states is slow, and the implied time scales are large.

In all three rows of Figure 7, the Markovian regime is reached much earlier for the implicit solvent model than for the explicit solvent model; i.e., the influence of the marginal degrees of freedom vanishes more quickly. This can be explained by the fact

that the set of marginal degrees of freedom is much smaller in transition matrices constructed from stochastic dynamics simulations. It only consists of the bond-angle degrees of freedom which equilibrate faster than the solvation shell in an explicit solvent model. Note that the emulation of the solvent by friction coefficients and stochastic kicks is by definition Markovian. In contrast to the simulations with an explicit solvent model, some of the curves in the right column of Figure 7 deviate already at small lag times from the constant regime, in particular those which correspond to small implied time scales in Table 2. Two reasons for this are conceivable: (i) similar to in the bit-flip model, transition matrices with small implied time scales approach the equilibrium matrix so closely that the numerical error of the eigenvalue calculation is not negligible or (ii) poorly sampled transitions become more and more dominating as the modes corresponding to the second and third eigenvectors decay and cause a divergence from Markovian behavior. Panel D shows that τ_{Markov} decreases if the friction coefficient increases. The left column of Figure 7 shows how τ_{Markov} changes if the density and the temperature are varied in simulations with an explicit solvent model. Analogously to the results for stochastic dynamics, τ_{Markov} decreases if the density increases (panels A and C). At high density, the kicks among solvent molecules and among solvent molecules and the solute are more frequent than at low density, leading to a quicker equilibration of the marginal degrees of freedom. Intuitively, a high density corresponds to a high value of the flipping probability p_E in the bit-flip model. Likewise, τ_{Markov} decreases if the temperature increases (panel E). For the higher temperatures, the kicks among solvent molecules do not necessarily become more frequent but have a higher impact, which also speeds up the equilibration of the solvent degrees of freedom.

5. CONCLUSION

We have presented an overview of the assumptions which are made when mapping the equations of motion onto the central quantity in Markov models, the transition matrix. We have also reviewed the mathematical properties of transition matrices. Markov models are a powerful tool to describe the dynamics of the relevant degrees of freedom of a system provided that one

finds a partition of the degrees of freedom of the system into relevant and marginal degrees of freedom such that the marginal degrees of freedom are not strongly coupled to the relevant degrees of freedom and that the former equilibrate on much shorter time scales than the latter. For liquid butane, we find that the discretization of the relevant degrees of freedom, if the grid boundaries do not mask the free energy barriers, has only little influence on the time resolution τ_{Markov} for which the dynamics becomes Markovian. The number of data points which are used to construct the Markov model, on the other hand, has an influence on the range of lag times for which the model is Markovian: the smaller the number of data points, the earlier the system diverges from Markovian behavior.

AUTHOR INFORMATION

Corresponding Author

*Phone: +41 44 632 5501. Fax: +41 44 632 1039. E-mail: wfvgn@igc.phys.chem.ethz.ch.

ACKNOWLEDGMENT

Financial support by the National Centre of Competence in Research (NCCR; Structural Biology) and by a grant (number 200021-121913) from the Swiss National Science Foundation (SNSF) and by a grant (number 228076) from the European Research Council (ERC) is gratefully acknowledged.

REFERENCES

- (1) van Gunsteren, W. F.; Berendsen, H. J. C. Algorithms for macromolecular dynamics and constraint dynamics. *Mol. Phys.* **1977**, *34*, 1311–1327.
- (2) Berendsen, H. J. C.; van Gunsteren, W. F. In *Molecular-Dynamics Simulation of Statistical-Mechanical Systems*, Proceedings of the International School of Physics “Enrico Fermi”, Varenna, Italy, 1985; Ciccotti, G., Hoover, W. H., Eds.; North-Holland: Amsterdam, 1986.
- (3) Chandler, D.; Berne, B. J. Role of constraints on the conformational structure of *n*-butane in liquid solvents - Comment. *J. Chem. Phys.* **1979**, *71*, 5386–5387.
- (4) van Gunsteren, W. F. Constrained dynamics of flexible molecules. *Mol. Phys.* **1980**, *40*, 1015–1019.
- (5) van Gunsteren, W. F.; Karplus, M. Effect of constraints on the dynamics of macromolecules. *Macromolecules* **1982**, *15*, 1528–1544.
- (6) van Kampen, N. G. *Stochastic Processes in Physics and Chemistry*, 2nd ed.; Elsevier Science Publishers B.V.: Amsterdam, The Netherlands, 1992.
- (7) Noé, F.; Horenko, I.; Schütte, C.; Smith, J. C. Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states. *J. Chem. Phys.* **2007**, *126*, 155102.
- (8) Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K. A.; Swope, W. C. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J. Chem. Phys.* **2007**, *126*, 155101.
- (9) Sarich, M.; Noé, F.; Schütte, C. On the approximation quality of Markov state models. *Multiscale Model. Simul.* **2010**, *8*, 1154.
- (10) Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Dill, K. A. Long-time protein folding dynamics from short-time molecular dynamics simulations. *Multiscale Model. Simul.* **2006**, *5*, 1214–1226.
- (11) Swope, W. C.; Pitera, J. W.; Suits, F. Describing protein folding kinetics by molecular dynamics simulations. 1. Theory. *J. Phys. Chem. B* **2004**, *108*, 6571–6581.
- (12) Swope, W. C.; Pitera, J. W.; Suits, F.; Pitman, M.; Eleftheriou, M.; Fitch, B. G.; Germain, R. S.; Rayshubski, A.; Ward, T. J. C.; Zhestkov, Y.; Zhou, R. Describing protein folding kinetics by molecular dynamics simulations. 2. Example applications to alanine dipeptide and beta-hairpin peptide. *J. Phys. Chem. B* **2004**, *108*, 6582–6594.
- (13) Muff, S.; Caflisch, A. Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a beta-sheet miniprotein. *Proteins: Struct. Funct. Bioinf.* **2008**, *70*, 1185–1195.
- (14) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 19011–19016.
- (15) Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. Progress and challenges in the automated construction of Markov state models for full protein systems. *J. Chem. Phys.* **2009**, *131*, 124101.
- (16) Vanden-Eijnden, E.; Venturoli, M. Markovian milestone with Voronoi tessellations. *J. Chem. Phys.* **2009**, *130*, 194101.
- (17) Keller, B.; Daura, X.; van Gunsteren, W. F. Comparing geometric and kinetic cluster algorithms for molecular simulation data. *J. Chem. Phys.* **2010**, *132*, 074110.
- (18) Deuffhard, P.; Huisinga, W.; Fischer, A.; Schütte, C. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. *Linear Algebra Appl.* **2000**, *315*, 39–59.
- (19) Deuffhard, P.; Weber, M. Robust Perron cluster analysis in conformation dynamics. *Linear Algebra Appl.* **2005**, *389*, 161–184.
- (20) Buchete, N. V.; Hummer, G. Coarse master equations for peptide folding dynamics. *J. Phys. Chem. B* **2008**, *112*, 6057–6069.
- (21) Pan, A. C.; Roux, B. Building Markov state models along pathways to determine free energies and rates of transitions. *J. Chem. Phys.* **2008**, *129*, 064107.
- (22) Buchete, N. V.; Hummer, G. Peptide folding kinetics from replica exchange molecular dynamics. *Phys. Rev. E* **2008**, *77*, 030902.
- (23) Muff, S.; Caflisch, A. ETNA: Equilibrium transitions network and Arrhenius equation for extracting folding kinetics from REMD simulations. *J. Phys. Chem. B* **2009**, *113*, 3218–3226.
- (24) Micheletti, C.; Bussi, G.; Laio, A. Optimal Langevin modeling of out-of-equilibrium molecular dynamics simulations. *J. Chem. Phys.* **2008**, *129*, 074105.
- (25) Bowman, G. R.; Ensign, D. L.; Pande, V. S. Enhanced Modeling via Network Theory: Adaptive Sampling of Markov State Models. *J. Chem. Theory Comput.* **2010**, *6*, 787–794.
- (26) Jensen, C. H.; Nerukh, D.; Glen, R. C. Sensitivity of peptide conformational dynamics on clustering of a classical molecular dynamics trajectory. *J. Chem. Phys.* **2008**, *128*, 115107.
- (27) Singhal, N.; Pande, V. S. Error analysis and efficient sampling in Markovian state models for molecular dynamics. *J. Chem. Phys.* **2005**, *123*, 204909.
- (28) Hinrichs, N. S.; Pande, V. S. Calculation of the distribution of eigenvalues and eigenvectors in Markovian state models for molecular dynamics. *J. Chem. Phys.* **2007**, *126*, 244101.
- (29) Noé, F. Probability distributions of molecular observables computed from Markov models. *J. Chem. Phys.* **2008**, *128*, 244103.
- (30) Zwanzig, R. Nonlinear generalized Langevin equations. *J. Stat. Phys.* **1973**, *9*, 215–220.
- (31) Frenkel, D.; Smit, B. *Understanding Molecular Simulation - From Algorithms to Applications*, 2nd ed.; Elsevier Academic Press: London, United Kingdom, 2002; Vol. 1 of Computational Science Series.
- (32) Schwabl, F. *Statistische Mechanik*, 3rd ed.; Springer-Verlag: Berlin Heidelberg New York, 2004.
- (33) MacCluer, C. R. The many proofs and applications of Perron's theorem. *SIAM Rev.* **2000**, *42*, 487–498.
- (34) Deuffhard, P.; Andreas, P. *Numerical Analysis in Modern Scientific Computing*, 2nd ed.; Springer-Verlag: Berlin Heidelberg New York, 2003; Vol. 1 of Texts in Applied Mathematics.
- (35) Christen, M.; Hünenberger, P. H.; Bakowies, D.; Baron, R.; Bürgi, R.; Gerke, D. P.; Heinz, T. N.; Kastenholz, M. A.; Kräutler, V.; Oostenbrink, C.; Peter, C.; Trzesniak, D.; van Gunsteren, W. F. The GROMOS software for biomolecular simulation: GROMOS05. *J. Comput. Chem.* **2005**, *26*, 1719–1751.
- (36) Hockney, R. W. The potential calculation and some applications. *Meth. Comp. Phys.* **1970**, *9*, 136–210.

(37) Schuler, L. D.; Daura, X.; van Gunsteren, W. F. An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. *J. Comput. Chem.* **2001**, *22*, 1205–1218.

(38) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of Cartesian equations of motion of a system with constraints - molecular-dynamics of *n*-alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.

(39) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Di Nola, A.; Haak, J. R. Molecular-dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

Atomic Velocity Projection Method: A New Analysis Method for Vibrational Spectra in Terms of Internal Coordinates for a Better Understanding of Zeolite Nanogrowth

Marc Van Houteghem,[†] Toon Verstraelen,[†] Dimitri Van Neck,[†] Christine Kirschhock,[‡] Johan A. Martens,[‡] Michel Waroquier,[†] and Veronique Van Speybroeck^{*,†}

[†]Center for Molecular Modeling, QCMM Alliance Ghent-Brussels, Ghent University, Technologiepark 903, B-9052 Zwijnaarde, Belgium

[‡]Center for Surface Chemistry and Catalysis, Leuven University, Kasteelpark Arenberg 23, B-3001 Heverlee, Belgium

 Supporting Information

ABSTRACT: An efficient protocol is presented to identify signals in vibrational spectra of silica oligomers based on theoretical molecular dynamics (MD) simulations. The method is based on the projection of the atomic velocity vectors on the tangential directions of the trajectories belonging to a predefined set of internal coordinates. In this way only contributions of atomic motions along these internal coordinates are taken into consideration. The new methodology is applied to the spectra of oligomers and rings, which play an important role in zeolite synthesis. A suitable selection of the relevant internal coordinates makes the protocol very efficient but relies on intuition and theoretical insight. The simulation data necessary to compute vibrational spectra of relevant silica species are obtained through MD using proper force fields. The new methodology—the so-called velocity projection method—makes a detailed analysis of vibrational spectra possible by establishing a one-to-one correspondence between a spectral signal and a proper internal coordinate. It offers valuable perspectives in understanding the elementary steps in silica organization during zeolite nanogrowth. The so-called velocity projection method is generally applicable on data obtained from all types of MD and is a highly valuable alternative to normal-mode analysis which has its limitations due to the presence of many local minima on the potential energy surface. In this work the method is exclusively applied to inelastic neutron scattering, but extension to the infrared power spectrum is apparent.

INTRODUCTION

Zeolites are microporous inorganic materials, mostly with aluminosilicate components, which exhibit crystal structures containing pores and cages large enough to permit the diffusion of small molecules. Zeolites are indispensable in many industrial applications, e.g., in heterogeneous catalysis, absorption and molecular separation, and ion exchange. These applications and the extension of their application field in many domains are a motivation for further investigation aiming at an even deeper insight into the behavior of a zeolite.^{1–3}

Understanding how zeolites nucleate and grow is of fundamental scientific and technological importance. Insight into the molecular mechanisms of structuring of silica can lead to the development of new hierarchical materials promising high potential for optimization of processes in catalysis and molecular separation. In addition, controlled zeolite synthesis could open new fields of application, such as optical electronics,⁴ bio-implants,⁵ etc. A variety of efforts have already been made to elucidate the early stages of zeolite growth.^{6–9} The formation of the siliceous zeolite silicalite with MFI topology is one of the best studied cases.^{10–20} Colloidal silicalite-1 is synthesized from hydrolysis of tetraethylorthosilicate (TEOS) as a monomeric silica source in aqueous tetrapropylammonium hydroxide (TPAOH) at room temperature.^{21–24} In this study we will focus on these early stages of zeolite formation from a modeling

perspective using vibrational spectroscopy: infrared (IR) and inelastic neutron scattering (INS) spectroscopy. These are important identification tools in zeolite synthesis, but the spectra are sometimes difficult to interpret. In ref 25 it was found that at the beginning of TEOS hydrolysis, small oligomers are formed which grow in number and size as the reaction progresses to form nanoparticles. Comparison between simulated and experimental IR patterns has illustrated how the silica contained in the colloidal nanoparticles evolve with time, leading from small five-ring oligomers toward successively more condensed five-ring species.^{22,24,26}

Two methods are conventionally used for the calculation of vibrational spectra linking spectral patterns with atomic motions: normal-mode analysis (NMA) based on static approaches and a Fourier-based technique which receives input from molecular dynamics (MD) simulations. NMA is computationally less demanding but is restricted to the harmonic approximation which allows only small deviations from a local or global minimum on the potential energy surface (PES). The restriction stems from the fact that NMA is in essence a second-order approximation of the minima in the PES. As a consequence, only one minimum can be calculated and just a small part of the

Received: September 20, 2010

Published: March 03, 2011

surface is taken into account, which limits the technique when multiple minima are present.²⁷ Static approaches can go beyond the harmonic approximation. We refer in particular to the work of Scribano and Benoit²⁸ where the vibrational self-consistent field method (VSCF) is combined with the single-to-all (STA) approach succeeding in the construction of a sparse PES at high ab initio levels of theory. With the second technique, data from MD simulations are used to calculate spectra. In the past many papers have already been published in literature aiming at identifying signals in the Raman or IR spectrum as signatures for specific vibrations characterizing the structure of the molecular system under study. It is not the intention to give a complete survey of all works published in this area, we limit ourselves to those related with vibrations in silica particles. In particular we refer to ref 29 as a pioneering work in which vibrational spectra have been interpreted by means of a normal-mode analysis in terms of symmetry coordinates predicted by point group theory. Other papers analyzed the vibrational eigenmodes with the help of MD to identify signals as signatures of ring structures in silica.³⁰ Their focus is mainly devoted to the description of symmetric stretching modes, such as breathing modes in rings, as these are recognized to be Raman active. The vibrational eigenmodes are projected onto the coherent breathing modes of the bridging oxygen atoms. They could in many cases provide an unambiguous interpretation for the origin of the Raman lines. Also the group of Smirnov has been very active in this domain.^{31–37} Power spectra belonging to a suitably chosen symmetric internal coordinate describing the breathing vibration of a specific ring are constructed and were found to be able to discriminate the various ring structures. To get a better insight into the participation of atoms of different types in the vibrational spectra, a power spectrum for each kind of atom was computed by Fourier transformation of the atomic velocity autocorrelation functions. Some MD techniques (e.g., ab initio MD) are computationally expensive while others are not (e.g., based on force fields). So it depends on the chosen MD method whether or not a large part of the PES can be scanned. The reliability of such numerical simulation methods to reproduce experimental results strongly depends on the method to calculate the PES. Accurate potentials are necessary to describe the atomic and molecular forces of the molecule under study and to reproduce the structural properties and dynamical data. In order to obtain a reliable spectrum it is essential to scan a large part of the PES and thus to simulate during a representative time frame (\sim ns range). In view of this, only force field-based methods are viable to calculate the PES as quantum mechanically based methods appear to be too expensive for larger systems and for longer time scales, although serious progress is made on this issue by the group of Parrinello^{38,39} by combining Car–Parrinello and Born–Oppenheimer MD. Most of the existing protocols to derive vibrational spectra from the data collected during MD simulations are applicable regardless of whether force fields or ab initio MD are used. Despite many attempts to interpret the vibrational spectra and to define the internal mode contributing to a certain band in the vibrational spectrum further model development remains desirable. Very recently Jacob and Reiher⁴⁰ developed a model in which the frequencies at which the bands in the vibrational spectra appear, and the total intensities of these bands can be interpreted in terms of localized modes. This method seems to be very appropriate in analyzing spectra of large molecules like polypeptides and proteins, but the protocol is constrained to static quantum chemical calculations.

In this paper we present a completely new methodology for the interpretation of vibrational spectra but based on molecular dynamical approaches. In the new model the atomic velocities are projected on properly selected directions fully determined by the internal mode from which we want to know its impact on the spectrum. Numerical applications of the new method are based on classical MD simulations. For the silica building blocks under investigation in this work, we use a silica-derived force field based on the gradient curves method⁴¹ which was previously successfully applied to investigate the MFI fingerprint in zeolites.²⁶

The infrared power spectrum is computed by taking the Fourier transform of the autocorrelation function of the time variation of the electric dipole moment. Similarly, the velocity power spectrum or INS spectrum is determined by the Fourier transform of the atomic velocities, which is easier to compute after a MD run as it only needs the instantaneous atomic velocities at each time step. Compared to NMA, this dynamical approach is more flexible in the treatment of nonharmonic problems, although also in static approaches progress is made in going beyond the harmonic approximation as already stated.

The first goal of this paper is to identify peaks in vibrational spectra of zeolite building blocks with specific internal degrees of freedom. In particular we were interested in how the size of rings and/or connectivity affects signals in the vibrational spectrum. For this purpose a new method has been developed enabling us to link the spectrum with the internal degrees of freedom of zeolite nanoparticles. This method provides a suitable tool that predicts which particular modes are influenced by changing size and topology of the silica building blocks. As such, it constitutes a significant progress in understanding the mechanisms during the early stages of the zeolite growth, and it gives complementary and valuable information in addition to the usual protocol of spectra analysis where experimentally measured spectra are compared with fully theoretical spectra extracted from simulations of the zeolite frameworks.^{34–37,42} The basic understandings of zeolite spectra and the lattice dynamics of zeolites are well-known.^{43–45} But a computational method for an unambiguous validation and confirmation of peak assignments in smaller silica particles can still be improved. In addition, it can in turn be helpful in the development and/or refinements of models to compute atomic forces, e.g., force fields. If a MD run does not yield the correct vibrational spectrum with regard to experiment, in the sense that some peaks are missing or not at the correct position, the new methodology makes it possible to specify the internal mode that is poorly described. Relevant terms in the force field can be adjusted accordingly to remove the discrepancies.

In ref 26 the MFI topology was studied in terms of shifts of peaks as MFI-structured nanoparticles grow. As the particles become bigger in size, the fingerprint band lowers in frequency from 650 to 550 cm^{-1} . An isolated five-membered ring (5T) reveals an IR-active vibration around 650 cm^{-1} . By connecting pentasil rings, slightly larger building blocks are constructed, but no substantial changes for the peak position have been observed. The situation changes if a more condensed structure is formed, such as the silica octamer, referred to as 8T (4×5) (see Figure 11); a sudden shift by approximately 50 cm^{-1} is observed. The red shift becomes even more substantial when the particles grow larger. This 550 cm^{-1} band is regarded as the spectroscopic

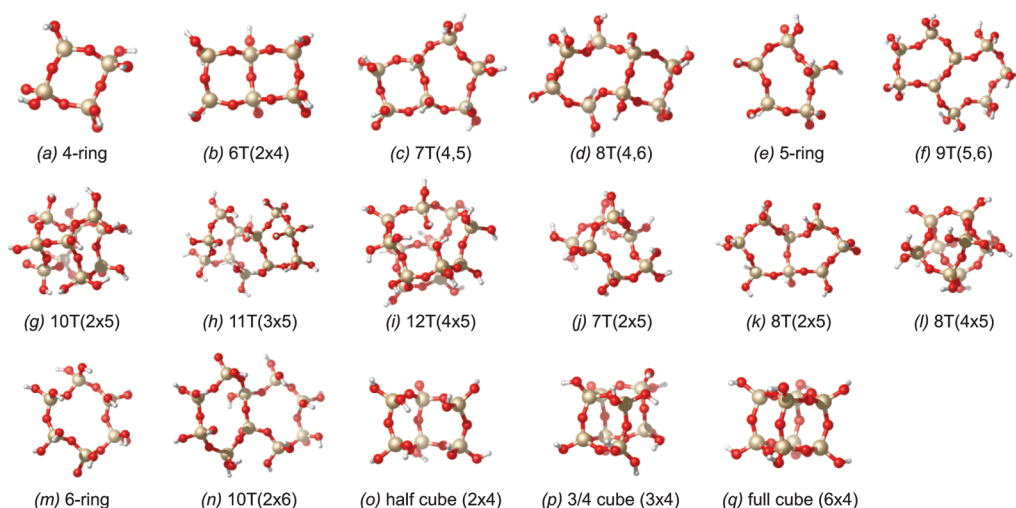


Figure 1. Simulated molecular structures of silica particles consisting of combinations between various four-, five-, and six-rings.

signature of MFI-type zeolites and is also a typical feature of similar structures built from five-membered rings, the so-called pentasil zeolites.

In this paper we apply a new methodology—the so-called velocity projection method—to a large variety of building blocks encountered in zeolite synthesis, including four-, five-, and six-rings and all types of connections between them (Figure 1). The analysis presented here should serve as a basic understanding of which modes and which peaks in the vibrational spectra are influenced by the particular interconnectivity and topology. The lack of experimental vibrational spectra of such small silica particles in the literature strengthens the importance of such a theoretical investigation. For larger species experimental data are available.

COMPUTATIONAL SECTION

The input of the computed spectra is generated from MD simulations. Interatomic interactions between framework atoms were represented by an in-house developed force field. It was especially designed for zeolites and was calibrated at the post-Hartree–Fock MP2/6-311+g(d,p) level of theory, with the gradient curves method (GCM).⁴¹ This is a novel technique that facilitates the development of transferable force-field models. It makes extensive use of regularization techniques⁴⁶ and of generic energy terms based on series expansions to obtain an optimal bias–variance trade-off during the fitting procedure. The force field was previously thoroughly benchmarked to reproduce the MFI fingerprint and IR-band shifts in other related small silica nanoparticles. All optimizations and simulations have been carried out with the CP2K⁴⁷ program package.²⁶

All initial geometries were built with the in-house developed software program package ZEOBUILDER.⁴⁸ Geometry optimizations were performed on all structures using the conjugate gradient method. With these optimized coordinates as input, an initial equilibration MD run of 5 ps was carried out. The actual production run which provided the data for analysis was an MD run of 1 ns at a temperature of 300 K with a Nose thermostat in the NVT statistical ensemble. To process the data, the MD-TRACKS program⁴⁹ was used.

From linear response theory, if the dipole moment history is obtained in an MD calculation, then the infrared spectrum $I_{\text{IR}}(\omega)$ can be computed as^{50–53}

$$I_{\text{IR}}(\omega) = \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{\alpha=x,y,z} \left| \int_0^t \frac{d\mu_{\alpha}}{dt} \exp(-i\omega t) dt \right|^2 \quad (1)$$

where ω is the angular frequency, and μ_{α} is the Cartesian components of the dipole moment. This is equivalent to the Fourier transform of the autocorrelation function of the time derivative of the dipole moment.^{53,54} The dipole moment is defined as

$$\boldsymbol{\mu} = - \int (\mathbf{r} - \mathbf{R}_c) \rho(\mathbf{r}) d\mathbf{r} + \sum_{i=1}^N Z_i (\mathbf{R}_i - \mathbf{R}_c) \quad (2)$$

where $\rho(\mathbf{r})$ is the electron density, Z_i represents the nuclear charge of atom i , \mathbf{R}_i is the position of atom i , and N the total number of atoms. The reference point \mathbf{R}_c is the center of charge for charged systems and is arbitrary for neutral systems. In molecular mechanics, one preferentially uses an effective charge Q_i for each atom, and the dipole moment is written as

$$\boldsymbol{\mu} = \sum_{i=1}^N Q_i (\mathbf{R}_i - \mathbf{R}_c) \quad (3a)$$

$$\frac{d\boldsymbol{\mu}}{dt} = \sum_{i=1}^N Q_i \mathbf{v}_i \quad (3b)$$

In this work fixed charges were used. The velocity power spectrum or inelastic neutron scattering spectrum (INS) can be obtained in a similar way as the infrared signal using an autocorrelation function:

$$I_{\text{INS}}(\omega) = \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{\alpha=x,y,z} \left| \int_0^t \mathbf{v}_{i,\alpha}(t) \exp(-i\omega t) dt \right|^2 \quad (4)$$

The IR and INS spectra are based on a Fourier transform which transforms information from the time domain to the frequency domain. One can easily see that both spectra are directly dependent on the velocities of the atoms $\mathbf{R}_{i,\alpha} = \mathbf{v}_{i,\alpha}$ so that the positions of peaks in IR and INS spectra coincide, but the intensities will vary.

The Fourier transform of the velocity autocorrelation provides information about the density of vibrational modes as a function of energy and hence reveals the underlying frequencies of the molecular system. The IR spectrum shows which internal modes are infrared active. If some of the IR active modes are localized on specific functional groups, then the corresponding bands can then serve as a spectral fingerprint of the groups. However, we focus in this paper on the INS spectra since charges in force fields are not (yet) well-defined in order to obtain reliable qualitative spectral amplitudes. Moreover, cancellation effects frequently appear in the time-varying dipole moment vector (eq 3a). It implies that the amplitudes in an IR spectrum are smaller and faster affected by relative noise. On the contrary, the INS spectrum contains only a summation of positive contributions (eq 4), and the relative noise on the spectrum becomes smaller. It also provides information about all vibrations of the system, i.e., all IR and Raman active modes.

METHOD

Velocity Projection Method. In order to associate peaks in the full INS spectrum to particular motions of internal coordinates we developed a method where the atomic velocities are projected to selective directions defined by the specific nature of the internal coordinate. An outline of this projection method is sketched below.

We make use of internal coordinates in order to assign a peak to one particular degree of freedom. An internal coordinate (IC) defines the location of the atoms in a molecule relative to the other atoms in the molecule. Some examples of IC's are bond distances, bending angles, dihedrals, Urey–Bradley stretches,⁵⁵ a linear combination of stretches, etc. Each internal coordinate q can be expressed as a function of the atomic Cartesian coordinates $\mathbf{R}_{i,\alpha}$ and depends on time: $q = q(\mathbf{R}_{i,\alpha}(t))$. The velocity projection method then consists of projecting the Cartesian velocity vector $\mathbf{v}_{i,\alpha} = \dot{\mathbf{R}}_{i,\alpha}$ of each atom i on the atomic tangent vector of the internal coordinate q at each time step. The IC's themselves are overcomplete and are, in contrast with NMA, not orthogonal to each other. In this section, the different classes of IC's that were used in this work will be illustrated. If one considers only one IC, then the definition of the tangent vector of atom i at each time step is given by

$$\mathbf{J}_{i\alpha} = \frac{\partial q}{\partial \mathbf{R}_{i,\alpha}} \quad (5)$$

where α runs over the Cartesian x , y , and z components of atom i . Obviously, i must be one of the atoms that specifies the IC $q(\mathbf{R}_{i,\alpha})$. The velocity vector belonging to atom i can be decomposed into its tangential and normal component with respect to the atomic tangent vector $\mathbf{J}_{i\alpha}$:

$$\mathbf{v}_{i\alpha} = (\mathbf{v}_{\perp})_{i\alpha} + (\mathbf{v}_{\parallel})_{i\alpha} \quad (6)$$

Only the tangential component gives a nonvanishing contribution to the change of the internal coordinate:

$$\delta q = \sum_i \mathbf{J}_i \cdot \delta \mathbf{R}_i \quad (7)$$

and with a suitable normalization can be set as

$$(\mathbf{v}_{\parallel})_{i\alpha} = \frac{(\sum_{j\beta} \mathbf{v}_{j\beta} \mathbf{J}_{j\beta}) \mathbf{J}_{i\alpha}}{\sum_{j\beta} \mathbf{J}_{j\beta}^2} \quad (8)$$

while the normal component lets the IC unchanged in first order (for small time steps):

$$(\mathbf{v}_{\perp})_{i\alpha} = \mathbf{v}_{i\alpha} - (\mathbf{v}_{\parallel})_{i\alpha} \quad (9)$$

The parallel velocity is the component that should be identified with this particular internal coordinate. Note that the IC is a function of time, and we project on a time dependent vector. To compute the actual spectrum, the Cartesian components of the velocities of every atom at each time step are plugged into eqs 1 and 4. The fluctuations due to the time dependency of the IC's are also computed with the Fourier transformation but should be small enough to neglect. Accordingly, an autocorrelation function (eq 4) is computed with the instantaneous $(\mathbf{v}_{\parallel})_{i\alpha}$ velocity projections. In this way a partial INS spectrum has been constructed restricted to those vibrations inducing changes of the IC of interest. A comparison between the full and partial INS spectra makes a full analysis of the vibrational spectrum possible: spectral peaks can be linked to IC's and corresponding (internal) degrees of freedom. When a peak in the projected spectrum coincides with its counterpart in the total spectrum, one can assume that there is no motion along other orthogonal coordinates that contributes to this peak. Such a peak is then completely resolved.

We note that the normalization of the tangential component (eq 8) is not unique and that a separate projection of the atomic velocity to the tangent vector belonging to that atom leads to a somewhat different normalization factor:

$$(\mathbf{v}_{\parallel})_{i\alpha} = \frac{(\sum_{j\beta} \mathbf{v}_{j\beta} \mathbf{J}_{j\beta}) \mathbf{J}_{i\alpha}}{\sum_{j\beta} \mathbf{J}_{j\beta}^2} \quad (10)$$

For computational reasons we have given preference to the normalization in (eq 8) to avoid numerical inaccuracies generated by tangent vectors \mathbf{J}_i of very small amplitude, as would be the case with the choice of solution 10.

Internal Coordinates. Analysis Using Internal Coordinates. At this point it is useful to outline the exact workflow. The standard procedure to obtain a vibrational spectrum from MD data is to compute the Fourier transform of the Cartesian atomic velocities (eqs 1 and 4) at each time step. In this work we project the Cartesian atomic velocities onto IC's to obtain the velocity component $(\mathbf{v}_{\parallel})_{i\alpha}$ of each atom alongside the IC (eq 8 and eq 9). Of these projected velocities, which are by definition equal to or smaller than the atomic Cartesian velocities, the Fourier transform is also computed at each time step to obtain the spectrum of the velocities for each IC separately. It can happen that, due to symmetry, some classes of IC's generate almost identical spectra; then the average was taken over them. In the next section the different classes of internal coordinates that were used in this work are discussed.

Classes for Internal Coordinates. The peaks in the full spectra can be (partially) decomposed by a proper selection of IC's. This can be done for the various zeolite building blocks under study in

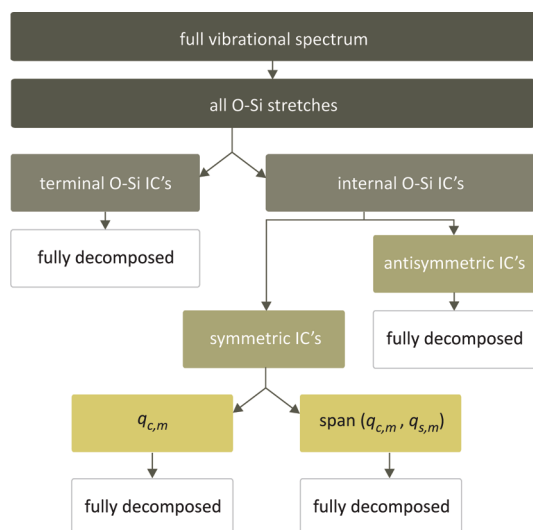


Figure 2. Flowscheme to classify different IC's used to decompose vibrational zeolite spectra.

this work and displayed in Figure 1. The IC's or linear combinations of them are then classified. Each class has its specific internal motion and generates a partial INS spectrum for each building block.

We will follow the following protocol. For all structures we compute the full vibrational spectrum based on the (nonprojected) atomic velocities as resulting from the MD simulation. For each individual IC we then project the atomic velocities according to the outlined procedure trying to link spectral peaks to these IC's. Linear combinations of various IC's can be constructed to form new IC's. With a well-chosen linear combination, one is able to fully assign a peak to this new specific internal coordinate. The general workflow for combining IC's to further unravel the spectra and to be able to link them with a spectral peak is shown in Figure 2. In this paper we focus on IC's constructed from stretches between two atoms: the stretches themselves can serve as IC's or linear combinations of them. The stretching vibration is one of the most straightforward internal motions to describe. In an earlier study by van Santen et al.,⁴⁵ it was stated that the lattice dynamics of amorphous silica can be correctly described by ignoring the Si–O–Si bending and solely focusing on the Si–O stretchings (see Figure 3). It implies that the wavenumber regime of 500–1200 cm⁻¹ is sufficient for studying the zeolite nanogrowth. As the studied species are however much smaller, we will validate in how far only stretches are applicable for fully resolving the spectrum. The higher the connectivity, the higher the rigidity and thus preventing the bending mode. Mixing of the Si–O stretching and Si–O–Si bending modes in the case of small oligomers may be assumed to be of minor importance because their modes are spectrally well separated.

Looking at Figure 3 we can discriminate between two types of O–Si stretches. The first one is the stretch mode, which occurs along the O–Si bond, where the oxygen atom is connected with a terminal hydrogen atom. This will be referred to as the O–Si terminal stretch. The second type consists of the two O–Si stretches in the Si–O–Si bridge. Such O–Si stretch is called an internal O–Si stretch. We make an approximation in the sense that the interaction between the symmetric and antisymmetric modes can be neglected, which is quite reasonable in view of the

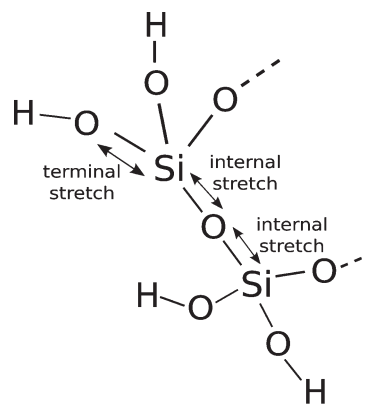


Figure 3. Schematic representation of the Si–O stretches present in the studied zeolite structures.

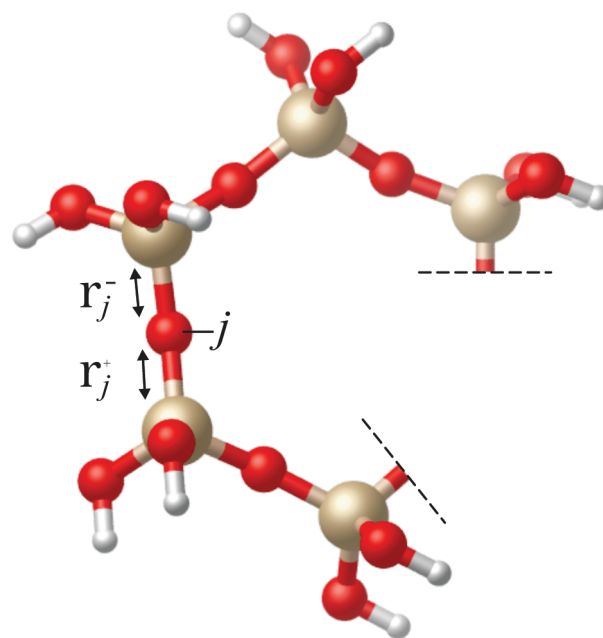


Figure 4. Ring stretching Si–O bonds defining symmetric and anti-symmetric IC's.

large difference between the spectral values of these modes and the magnitude of the coupling constant.⁴⁵

One can choose a particular Si–O stretch as IC, but in practice, linear combinations of stretch modes are more appropriate choices for IC's. In this way we separate the $2N$ ring stretching Si–O bonds, present in a N -membered ring, into two classes of IC's (symmetric and antisymmetric). For one particular Si–O–Si bridge j :

$$q_{ss,j} = \frac{1}{\sqrt{2}}(r_j^- + r_j^+) \quad (11)$$

and

$$q_{as,j} = \frac{1}{\sqrt{2}}(r_j^- - r_j^+) \quad (12)$$

as visualized in Figure 4.

The coordinate transformation between all stretching internal coordinates and the N symmetric and antisymmetric IC's is orthogonal and of dimension $2N \times 2N$:

$$\begin{bmatrix} \mathbf{q}_{ss} \\ \mathbf{q}_{as} \end{bmatrix} = \mathbf{T} \mathbf{r} \quad (13)$$

$\mathbf{r} = (2N \times 1)$ matrix (Si – O stretches)

$\mathbf{T} = 2N \times 2N$ matrix and orthogonal

The force constant matrix in $\{\mathbf{q}_{ss}, \mathbf{q}_{as}\}$ space becomes

$$\mathbf{U} = \frac{1}{2} (\mathbf{q}_{ss}^T, \mathbf{q}_{as}^T) \mathbf{C}_q \begin{bmatrix} \mathbf{q}_{ss} \\ \mathbf{q}_{as} \end{bmatrix} \quad (14)$$

with $\mathbf{C}_q = \mathbf{T} \mathbf{C}_r \mathbf{T}^T$, which is not block diagonal even with a diagonal structure of \mathbf{C}_r :

$$\mathbf{C}_q = \begin{bmatrix} \mathbf{C}_{q_{ss}} & \mathbf{C}_{q_{coupl}} \\ \mathbf{C}_{q_{coupl}} & \mathbf{C}_{q_{as}} \end{bmatrix} \quad (15)$$

In the model proposed it is assumed that the coupling between the symmetric and antisymmetric stretch modes is very weak and may be omitted: $\mathbf{C}_{q_{coupl}} = 0$. This seems to be justified since the frequency bands of both classes are well separated, as will be demonstrated later in the discussion section.

It is interesting to examine the symmetric stretch mode in more detail. The peaks due to the symmetric stretch are not completely resolved, and they lie in the wavenumber region which is interesting for obtaining insight into zeolite synthesis (the antisymmetric stretch peaks are completely resolved and lie in a less interesting area). In order to gain insight into the normal (symmetric stretch) modes of a molecular ring system, we consider a simplified model of n identical and equally spaced masses on a circle. Due to its periodicity, the normal modes of such a system can be obtained as the eigenvectors of a so-called circulant matrix C :

$$C_{jk} = c_{(j-k) \bmod n} \quad (16)$$

for $j, k = 1, \dots, n$, where the notation $(i) \bmod n$ implies integer arithmetic modulo n . Since the Hessian of the model system is a symmetric matrix ($C_{jk} = C_{kj}$), the coefficients must also obey

$$c_{(i) \bmod n} = c_{(-i) \bmod n} \quad (17)$$

where we used the notation $-i = n - i$. The circulant matrix C for an even n then has the form:

$$C = \begin{bmatrix} c_0 & c_1 & \dots & c_{n/2} & c_{n/2-1} & \dots & c_1 \\ c_1 & c_0 & \dots & & & & \\ \vdots & & \ddots & & & & \\ c_{n/2} & & & \ddots & & & \\ c_{n/2-1} & & & & \ddots & & \\ \vdots & & & & & \ddots & \\ c_1 & & \dots & & & & c_1 & c_0 \end{bmatrix} \quad (18)$$

The form of C for n being odd is analogous. In the remainder we drop the explicit "mod n " notation for the $c_{(i)}$ indices, but it is understood that they should be interpreted in arithmetic modulo n . The eigenvectors of the circulant matrix can be used to create new IC's on which we can project the atomic velocities, based on

the symmetric stretch IC's $q_{ss,k}$. With the circulant matrix being the model for the Hessian, we can construct these new internal coordinates, i.e., linear combinations of symmetric stretches, to further resolve the spectra of silica rings.

The eigenvalues and the k^{th} component of the normalized eigenvectors are (see Appendix Section):

$$\begin{aligned} \lambda^{(m)} &= \sum_k c_{(k)} \exp\left(\frac{-2\pi i m k}{n}\right) \\ &= \sum_k c_{(k)} \cos\left(\frac{2\pi m k}{n}\right) = \lambda^{(-m)} \end{aligned} \quad (19)$$

and

$$V_k^{(m)} = \frac{1}{\sqrt{n}} \exp\left(\frac{2\pi i m k}{n}\right) \quad (20)$$

Looking at the eigenvalue spectrum one can distinguish between even n (four-, six-ring, etc.) and odd n (five-ring, etc.). For even n the eigenvalues (eigenvectors) become

$$\begin{cases} \lambda^{(0)}(V^{(0)}) : \text{nondegenerate} \\ \lambda^{(m)}(V^{(m)}, V^{(-m)}) \text{ for } m = 1, \dots, \frac{n}{2} - 1 : \text{two-fold degenerate} \\ \lambda^{(n/2)}(V^{(n/2)}) : \text{nondegenerate} \end{cases}$$

For odd n they are

$$\begin{cases} \lambda^{(0)}(V^{(0)}) : \text{nondegenerate} \\ \lambda^{(m)}(V^{(m)}, V^{(-m)}) \text{ for } m = 1, \dots, \frac{n-1}{2} : \text{two-fold degenerate} \end{cases}$$

It is practical to switch to real eigenvectors for the two-fold degenerate eigenspaces, by taking linear combinations:

$$W_k^{(m)} = \frac{1}{2} (V_k^{(m)} + V_k^{(-m)}) = \cos\left(\frac{2\pi}{n} m k\right) \quad (21a)$$

$$W_k^{(-m)} = \frac{1}{2i} (V_k^{(m)} - V_k^{(-m)}) = \sin\left(\frac{2\pi}{n} m k\right) \quad (21b)$$

It is now possible to construct a new internal coordinate for each eigenvector m by making linear combinations of the symmetric stretches $q_{ss,k}$ where the coefficients are the elements of the eigenvector. For the circular systems under consideration, the new internal coordinates become

$$\begin{aligned} q_{c,m} &= \sum_k W_k^{(m)} q_{ss,k} = \sum_k \frac{V_k^{(m)} + V_k^{(-m)}}{2} q_{ss,k} \\ &= \sum_k \cos\left(\frac{2\pi}{n} m k\right) q_{ss,k} \end{aligned} \quad (22a)$$

$$\begin{aligned} q_{s,m} &= \sum_k W_k^{(-m)} q_{ss,k} = \sum_k \frac{V_k^{(m)} - V_k^{(-m)}}{2i} q_{ss,k} \\ &= \sum_k \sin\left(\frac{2\pi}{n} m k\right) q_{ss,k} \end{aligned} \quad (22b)$$

where n is the number of symmetric stretches in the circular molecular system and $q_{ss,k}$ is short for the k^{th} symmetric stretch IC. Since $q_{c,m}$ and $q_{s,m}$ only differ in phase we could not discriminate between them. Therefore it is useful to consider the subspace spanned by these vectors. This subspace is referred to as $\text{span}(q_{c,m}q_{s,m})$. At each time step of the simulation we project the atomic velocities on the set of vectors which spans this subspace. For topologically symmetric systems the INS spectra of $q_{c,m}$, $q_{s,m}$ and $\text{span}(q_{c,m}q_{s,m})$ coincide (e.g., Figure 1b, n, and q and a, e, and m).

Note that this technique can only be used for single closed ring systems, and it is not expected to completely resolve spectra in systems with fused rings. If multiple rings are attached to each other, then one has to distinguish between them to make use of the circulant matrix as a model for the Hessian. This is still useful because the resulting spectra can be directly compared with spectra from other n -ring structures. In this way the immediate effect of a different topology on a specific structure can be studied.

RESULTS AND DISCUSSION

Figure 1 gives an overview of the three-dimensional (3D) optimized geometries of the simulated structures. Some of them are key components of the MFI structure, and others have been suggested to occur in the early stages of other zeolite structures.²⁶ The initial geometry as input for the MD simulation does not need to coincide with the global minimum as all possible structures on the PES will be visited during the MD run. The most elementary structures are a single four-, five- and six-membered ring (Figure 1a, e, and m), and their spectra may be regarded as reference. Larger building blocks can be constructed by connecting these elementary rings with other n -ring structures. These new structures yield new (partial) INS spectra which can be compared with the reference spectra, and such an investigation provides us information on the influence of the topology (i.e., how atoms are linked to each other) on the vibrational spectra. In the synthesis of zeolites with MFI topology it has been found that the five-membered ring (5T) (Figure 1e), also called the pentasil ring, is of special importance.^{56–58} By connecting these pentasil rings larger building blocks can be constructed. If three T-atoms are added, then two 8T(2 × 5) (Figure 1k) and three 11T(3 × 5) (Figure 1h) sideways five-membered rings arise. If two sides of two different five rings are shared (sharing three T-atoms), then the 7T five ring is formed, 7T(2 × 5) (Figure 1j). The five ring 8T(4 × 5) (Figure 1l) is a silica octamer, where the rings form a cagelike structure. The species where two five rings do not share but are connected with an oxygen bridge is referred to as 10T(2 × 5) (Figure 1g). In this topology an extra six ring is present. Finally the silica dodecamer structure forms the 12T(4 × 5) (Figure 1i).²⁶

In aqueous silicate solutions the cubic octamer^{59,60} often has been observed experimentally and proposed as a building block. Therefore we also consider this full cube (6 × 4) (Figure 1q), a half cube (2 × 4) (Figure 1o), and a three-fourths cube (3 × 4) (Figure 1p).

Physical Meaning of IC's. The most basic molecular structures are the four-, five-, and six-rings (Figure 1a, e, and m, respectively). Their spectra may serve as a reference for other topologies where n_1 -rings are attached to n_2 -rings. In this way, the vibrational spectra can be applied as a tool for studying topological differences in zeolite structures (e.g., during nanogrowth).

Looking at the diagram of Figure 2, we can distinguish between two types of IC's: those that can be identified as single stretches and those that show a weighted mixture of stretch modes. For the first type of IC's it is straightforward to assign a physical motion to them since they coincide with real stretches. For the second type of IC's it is less straightforward, but they also represent a real physical combined vibration.

Symmetric and antisymmetric ring vibrations have been defined in eqs 11 and 12. When the atomic velocities are projected on these IC's, we gain knowledge on the magnitude of these combined modes relative to the full vibrational spectrum.

The antisymmetric motions represent modes where the O atom oscillates between two Si atoms. Their spectral lines are located at frequencies larger than 800 cm^{-1} and are discussed in subsection Classes for Internal Coordinates.

The characteristic internal modes of a molecular system are determined by the eigenvalue equation of the Hessian:

$$HE = \omega^2 ME \quad (23)$$

Here H is the Hessian which contains the second derivatives of the potential energy with respect to the Cartesian coordinates, M is the mass matrix, ω^2 represents the eigenvalues, and E represents the $3N \times 3N$ matrix which contains the eigenvectors of the different modes (N = number of atoms). As shown in the Internal Coordinates Section, we introduced the circulant matrix as a model for the Hessian corresponding to the symmetric IC's and determined its eigenvectors and eigenvalues. The physical meaning of the combined modes of IC's $q_{c,m}$ and $q_{s,m}$ is interesting for further survey. For the four-, five-, and six-ring we have the following IC's (see eqs 22a and 22b) ($\alpha = 2\pi/n$):

- Four-ring: $q_{c,0}$, $\text{span}(q_{c,1}q_{s,1})$, $q_{c,2}$ ($\alpha = 90^\circ$).
- Five-ring: $q_{c,0}$, $\text{span}(q_{c,1}q_{s,1})$, $\text{span}(q_{c,2}q_{s,2})$ ($\alpha = 72^\circ$).
- Six-ring: $q_{c,0}$, $\text{span}(q_{c,1}q_{s,1})$, $\text{span}(q_{c,2}q_{s,2})$, $q_{c,3}$ ($\alpha = 60^\circ$).

They all can be associated to a real physical mode. Consider, for example, the four-ring. When projecting on the corresponding IC's, the resulting spectral peaks agree with the molecular breathing motions, as depicted in Figure 5, in which an outward arrow means that the particular bond stretches out, whereas an inward arrow points to a shrinking motion of the bond. From this figure it is clear that $q_{c,1}$ and $q_{s,1}$ actually represent the same breathing mode, only shifted in phase by 90° . Thus, whenever a peak arises when projecting on a particular IC, we know with which vibrational eigenmode it corresponds. As a special combined mode we define in this work the breathing mode, in which all the present Si–O stretches are elongated or shrunk simultaneously. Here the mode $q_{c,0}$ represents this breathing mode.

Elementary Building Blocks: Four-, Five-, and Six-Rings.

The spectra of all structures considered in this paper for all IC's are taken up in the Supporting Information. In this section some relevant spectra of the four-, five-, and six-rings are shown and will be further explained. We will focus on the spectral shifts that occur when spectra of the four-, five-, and six-rings for the same type of IC are compared. The conclusions that are drawn from these elementary structures are transferable to more complex molecular structures. Also for the sake of completeness INS and IR spectra for the four-, five-, and six-membered ring systems are given in the Supporting Information. They can serve as comparative material when analyzing velocity power spectra associated to specific IC's.

Terminal O–Si and Antisymmetric IC's. The spectra of the terminal O–Si and antisymmetric IC's are shown in Figure 6 for

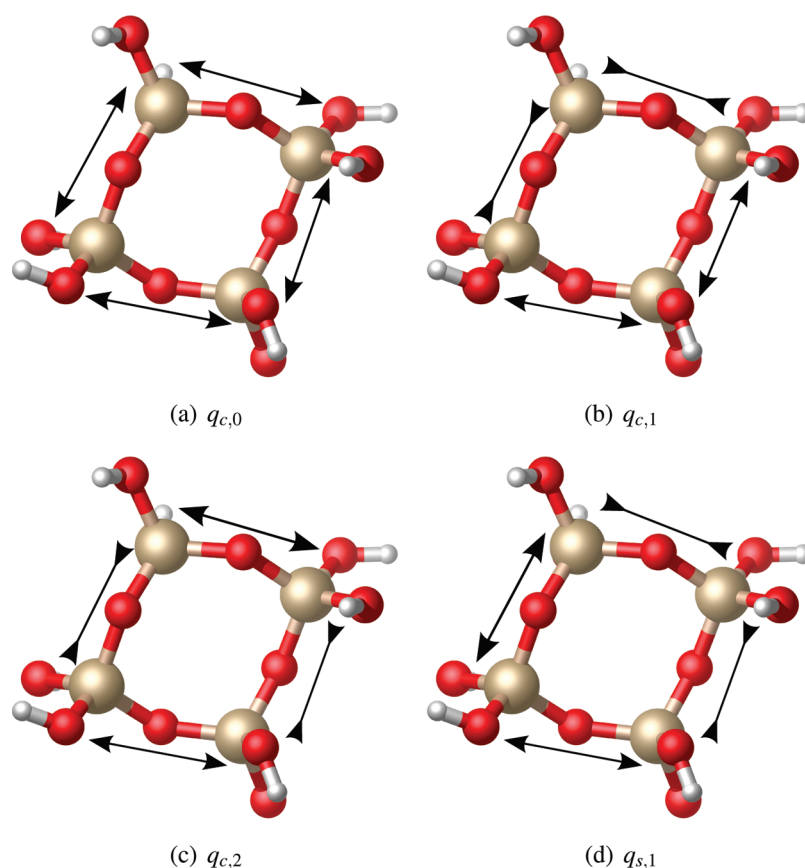


Figure 5. Visualization of the physical breathing modes when projecting atomic velocities on eigenvectors of the circulant matrix as a model for the Hessian.

the three elementary structures. For these classes of stretches multiple IC's are present. Per category the average spectrum is taken, as the spectra belonging to each IC of the same class are in essence not deviating much from each other. Also the standard deviation (stdev) of this average is given (lower curve). For both types of IC's the spectra manifestly show little influence on the size of the ring (same position, relative amplitudes, and shape).

Circulant Symmetric IC's. As outlined in the Method Section, the symmetric IC's are analyzed making use of the properties of the circulant matrix as a model for the Hessian. The spectra of the velocity projections on the $q_{c,m}$ and $q_{s,m}$ IC's of the four-, five-, and six-ring are shown in Figures 7 and 8. From these figures, some interesting features related to ring size can be derived.

- (i) Comparing the first eigenmode, $q_{c,0}$, of the four-, five-, and six-ring we see that the spectra of this breathing mode only exhibit a very small spectral shift which is almost negligible. The spectrum of the four-ring has peaks at 549 and 888 cm^{-1} , and for the five- and six-rings, we get 553 and 887 cm^{-1} and 562 cm^{-1} and 888 cm^{-1} , respectively. A small third peak emerges at about 453 cm^{-1} when increasing the size of the ring. It is at this stage not clear to which motion this third vibrational peak could be associated. The two peaks in the four-ring spectrum probably belong to the breathing modes of the ring Si and O atoms with a growing impact on the stretches. As the size of the ring increases, a collective nonplanar vibration could emerge. Although intuitive to some extent, these

explanations remain speculative as so far no visualization program is available which could remove this ambiguity.

- (ii) The peaks of the second mode of the four-ring, span($q_{c,1}$, $q_{s,1}$), make a shift compared to the same mode of the five- and six-rings. The two peaks with the four-ring occur at 642 and 869 cm^{-1} , the five- and six-rings show peaks at 617 and 876 cm^{-1} and 602 and 881 cm^{-1} , respectively. There is some peak drift, and the interpeak distance grows with increasing ring size (red shift for the left peak, blue shift of the right peak). They are still relatively small but big enough to be measurable and to conclude that there is a size dependence, which is more pronounced than in the $q_{c,0}$ mode.
- (iii) The above made conclusions are also valid when the eigenmode span($q_{c,2}$, $q_{s,2}$) of the five- and six-rings are compared.
- (iv) The interpretation of the spectra for these elementary building blocks, however, is also valid when combinations of these structures are made, as can be seen from the figures taken up in Figure 12 and the Supporting Information, e.g., 7T(4,5), 8T(4,6), and 9T(5,6). There is a very small topology dependence for the $q_{c,0}$ mode, and the influence of the ring size is larger for higher order modes.

We can discuss now how far the above features agree with what has been published in literature. The independence of spectra of the $q_{c,0}$ breathing coordinate on the ring size contradicts results of previous modeling studies. These works have demonstrated a

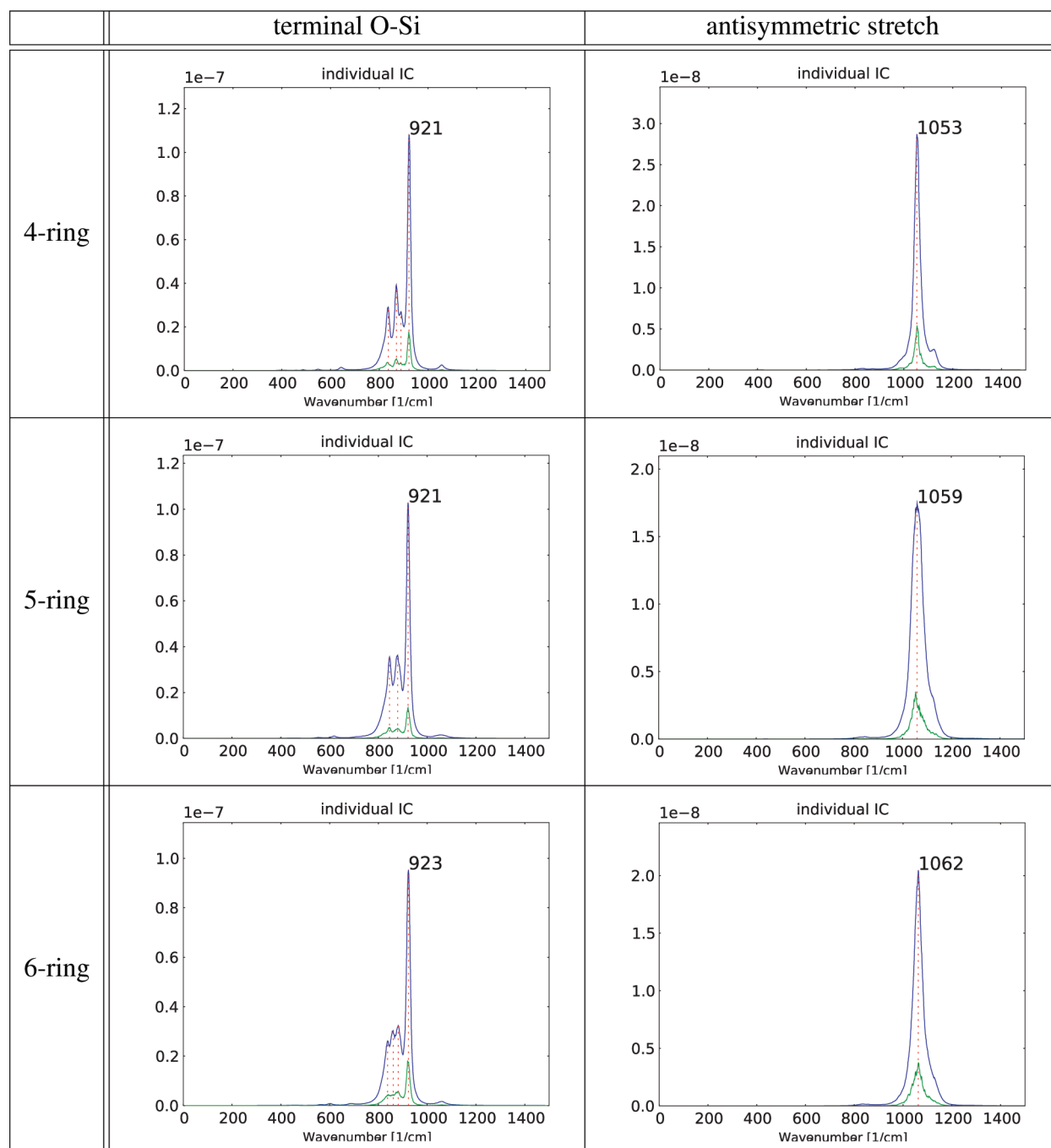


Figure 6. Projection of atomic velocities on the terminal O–Si and antisymmetric stretch IC's.

downward shift of the “ring breathing frequencies” with the increase of ring size. In the paper of Smirnov et al.³¹ the internal coordinate measuring the ring pore-opening has been put equal to

$$P(t) = \frac{1}{N/2} \sum_{i=1}^{N/2} R_i(t) \quad (24)$$

where $R_i(t)$ stands for the deviation of the i^{th} ring diameter from the mean value. The ring diameter is defined as the distance between oxygen atoms on the opposite sides of the ring. In order to comment on this issue, we have computed the velocity power spectrum of this vibrational coordinate $P(t)$ making use of the

MD results. These calculations could reveal the origin of the apparent discrepancies between the results presented in this work and those of literature. We also did the same exercise for the silica, instead of the oxygen atoms. The results are surprising and are displayed in Figure 9. Some peaks occurring in the $P(t)$ velocity power spectra for the silica atoms perfectly coincide with the $q_{c,0}$ in both four-, five-, and six-ring spectra. More specifically, four-ring: 550 and 892 cm^{-1} ; five-ring: 550 and 883 cm^{-1} ; and six-ring: 558 and 892 cm^{-1} . They do not show any ring size dependence. On the contrary, the oxygen $P(t)$ velocity power spectra do not show any resemblance with the $q_{c,0}$, except perhaps for the spectral signal at 453 cm^{-1} appearing in six-membered rings. The downward shift of the ring breathing

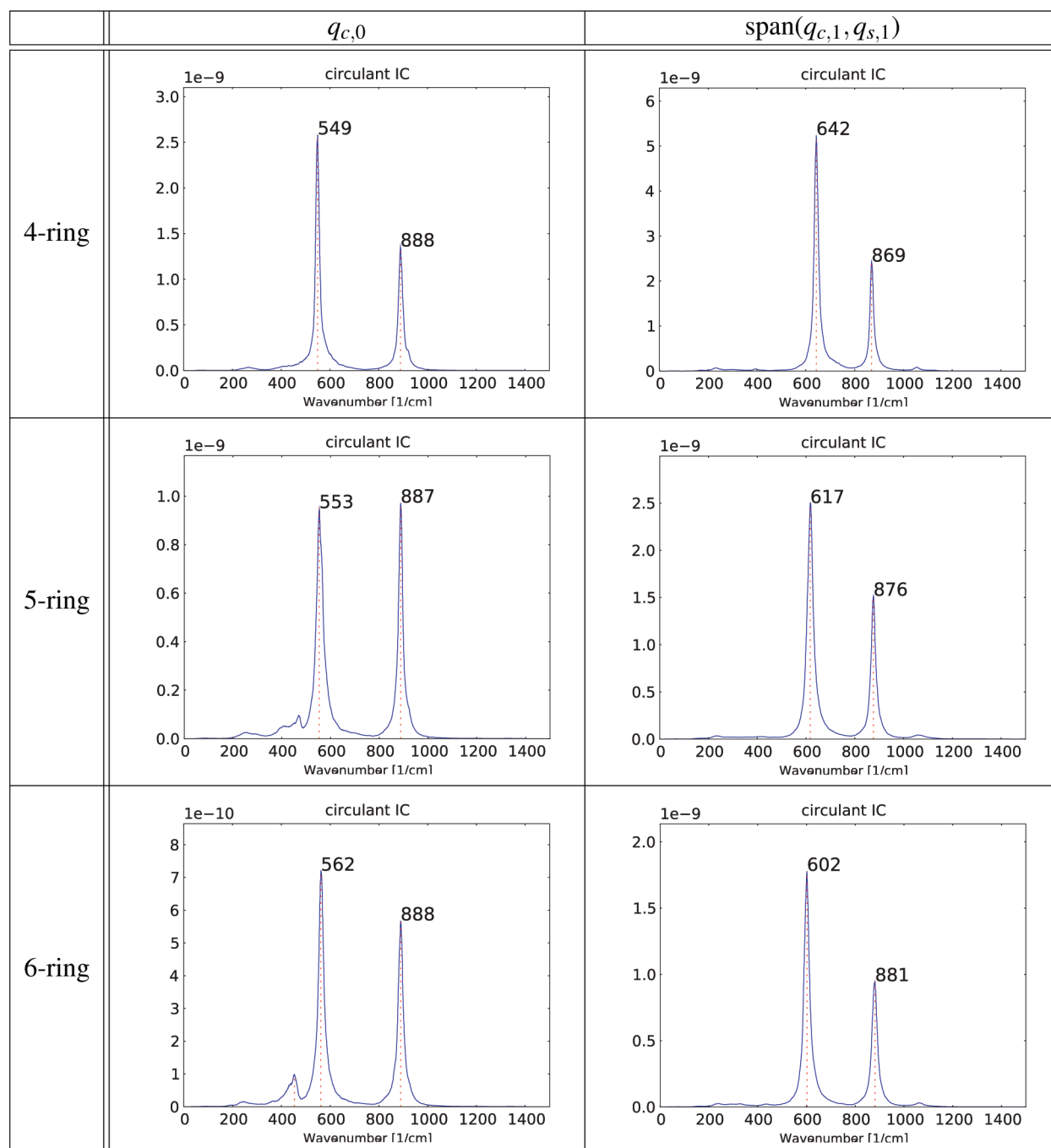


Figure 7. Projection of atomic velocities on the circulant symmetric IC's $q_{c,0}$ and $\text{span}(q_{c,1}, q_{s,1})$ for the four-, five-, and six-ring.

frequencies observed in literature is also reproduced in our $P(t)$ velocity power spectra for the oxygen atoms: 483, 475, and 458 cm^{-1} in going from four- to six-rings. They represent the lowest frequency peak in the spectrum. What can we conclude from this comparative study? Which peak corresponds really with the breathing mode where all ring diameter vibrations are supposed to be in phase? Since the pore-opening vibrational coordinate $P(t)$ does not take into account the phase of all ring diameter vibrations, the $P(t)$ velocity power spectra show admixtures of various ring modes, as shown in Figure 5 in case of a four-ring.

For completeness, we also introduced another internal coordinate being the sum of all O–O distances of neighboring oxygen bridges as suggested in ref 32. We adapted the velocity

projection method on this internal coordinate, and the resulting spectra are given in the Supporting Information. These spectra reveal a striking resemblance with previous $P(t)$ velocity power spectra both for the oxygen as for the silica atoms.

The presence of identical peaks in the three different types of power spectra (three different internal coordinates, namely $q_{c,0}$, $P(t)$ and $\sum R_{\text{O-O}}$) points manifestly toward a common vibrational mode, which represents probably the “true” breathing mode. However, to remove any ambiguity, we should be able to visualize the vibration corresponding to each peak in the power spectrum. But this is not an easy task and falls outside the scope of this work. Another aspect concerns the role of the force field used to perform the MD runs. It should be stressed that the fact that all

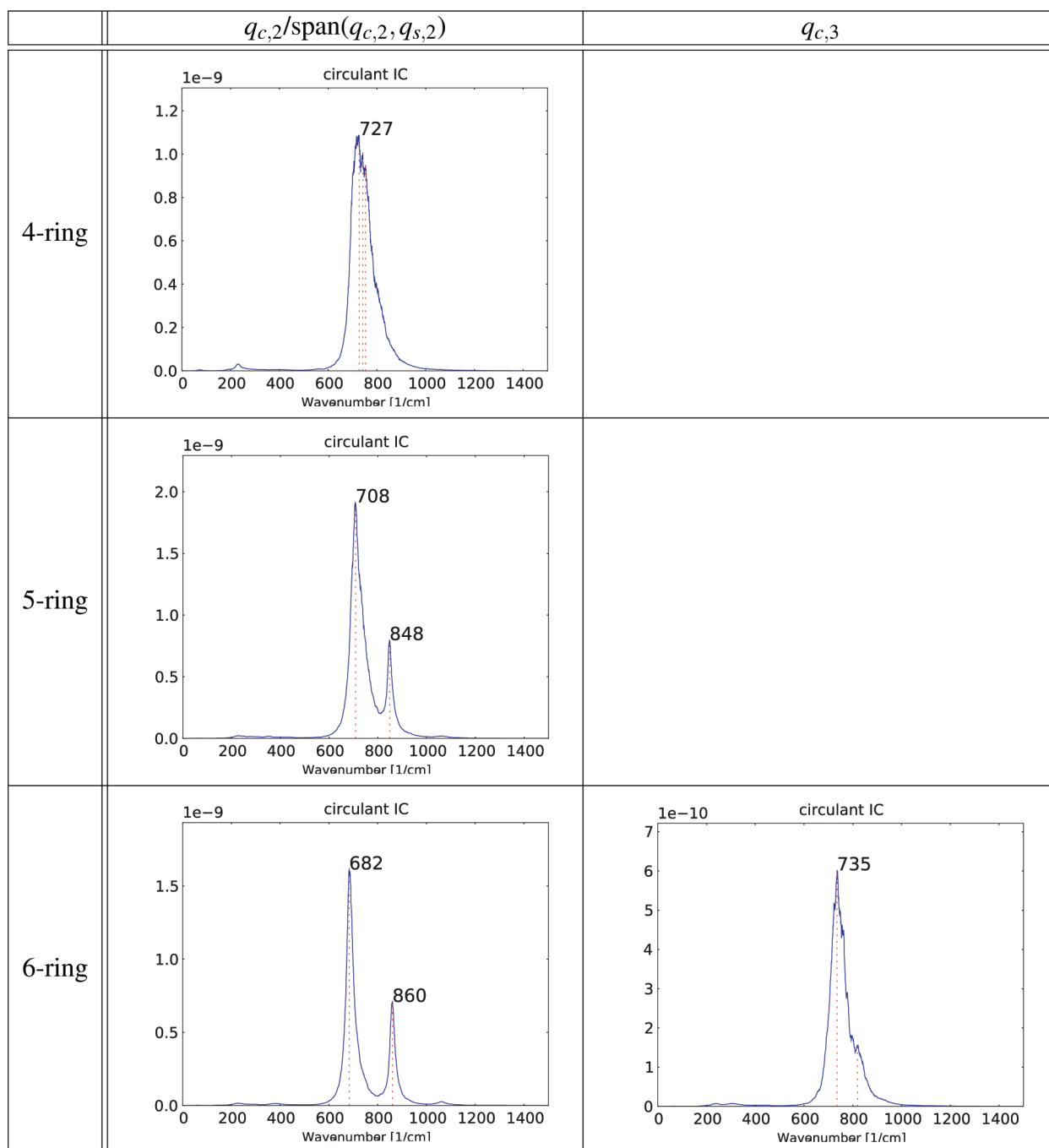


Figure 8. Projection of atomic velocities on the circulant symmetric IC's $q_{c,2}/\text{span}(q_{c,2}, q_{s,2})$ and $q_{c,3}$ for the four-, five-, and six-ring.

peaks observed in the $q_{c,0}$ spectra are also retrieved in the standard “non-projected” velocity power spectra corresponding with pore-opening IC's is a property independent of the choice of the force field. Use of another force field can reproduce the spectral lines at somewhat different wavenumbers and affects probably too the shift of a peak when increasing the ring size, but the general features, as sketched above, remain unaffected.

Projection on Orthogonal Basis of IC's and the Orthogonal Complement. In the previous section only one class of internal coordinates (O–Si stretches and linear combinations of them) has been investigated. In this section we want to see to what extent the restriction to this single class of stretches is accurate

enough. In other words what is the impact of the bending and dihedral motions on the vibrational spectra, or, more precisely, what is the impact of the remainder on the spectra after projecting out all tangent vectors belonging to stretches? The projection technique consists of determining all tangential atomic velocity vectors inducing a change in a particular IC. Modes induced by the normal components of the velocities are not considered as far as they do not belong to the entire class of IC's (stretches). For that reason we develop a method able to construct the so-called orthogonal complement of the basis of a given set of internal coordinates containing all the O–Si stretches (terminal O–Si + internal O–Si IC's). To obtain this orthogonal complement, the singular value decomposition of the

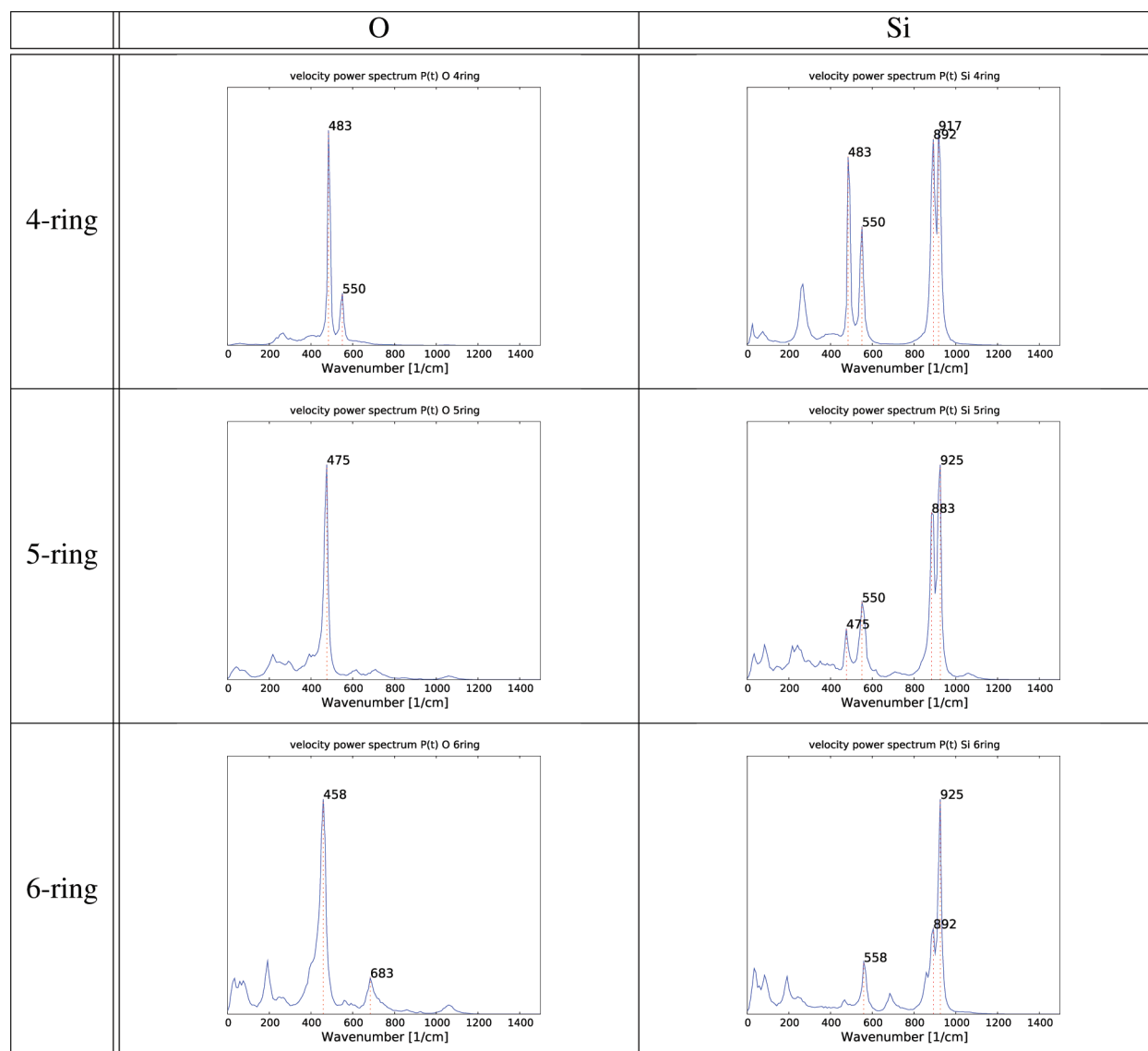


Figure 9. Velocity power spectra $P(t)$ for oxygen atoms as well as for silica and for the four-, five-, and six-rings.

matrix \mathbf{A} holding the tangent vectors of all the IC's is computed at every time step:

$$A_{ki} = \frac{\partial q_k}{\partial x_i} \quad (25)$$

where $k = 1, \dots, K$ is an index characterizing the IC's, and the subscript $i = 1, \dots, 3N$ denotes the Cartesian coordinates. This procedure decomposes \mathbf{A} into three matrices:⁶¹

$$\mathbf{A} = \mathbf{U}\mathbf{W}\mathbf{V}^T \quad (26)$$

The matrices \mathbf{U} and \mathbf{V} are orthogonal, and \mathbf{W} has singular values and is diagonal. The first K columns of \mathbf{V} form an orthogonal basis for the IC's at each time step, while columns K to $3N$ construct the basis for the orthogonal complement at each time step. In a second step the atomic velocities are projected on both orthogonal bases. The result for the four-, five-, and six-rings is shown in Figure 10. The two full INS spectra are clearly well separated. The green spectrum ranging from 500 to 1200 cm^{-1} covers all O–Si stretches. The blue spectrum involves all IC modes belonging to the orthogonal complement, and for the

three-ring structures, all peaks are located below 500 cm^{-1} . This is a nice result as it demonstrates that our methodology works quite well in projecting out all bending and dihedral motions. The conclusion is that the frequency region in question between 500–1200 cm^{-1} is adequately probed by stretch IC's and that a one-to-one comparison with the full spectrum is possible. Analysis of the lower frequency spectrum should be done with care. The blue spectrum in Figure 10 grouping all IC's, excluding stretches, is a global spectrum. We did not project on the tangential vector of an individual IC belonging to the class of bending and dihedral angles. In principle the same procedure as in the previous subsection could be performed, but in view of the ultimate goal of this study, such investigation was expected to give little added value. Nevertheless additional information was extracted, which could be of interest. To illustrate with an example, we refer to the peak at around 25 cm^{-1} which is prominently present in the low-frequency spectra of the three-ring systems. This peak is probably due to a collective puckering mode of the ring, but as already stressed, a full treatment of the velocity projection protocol on

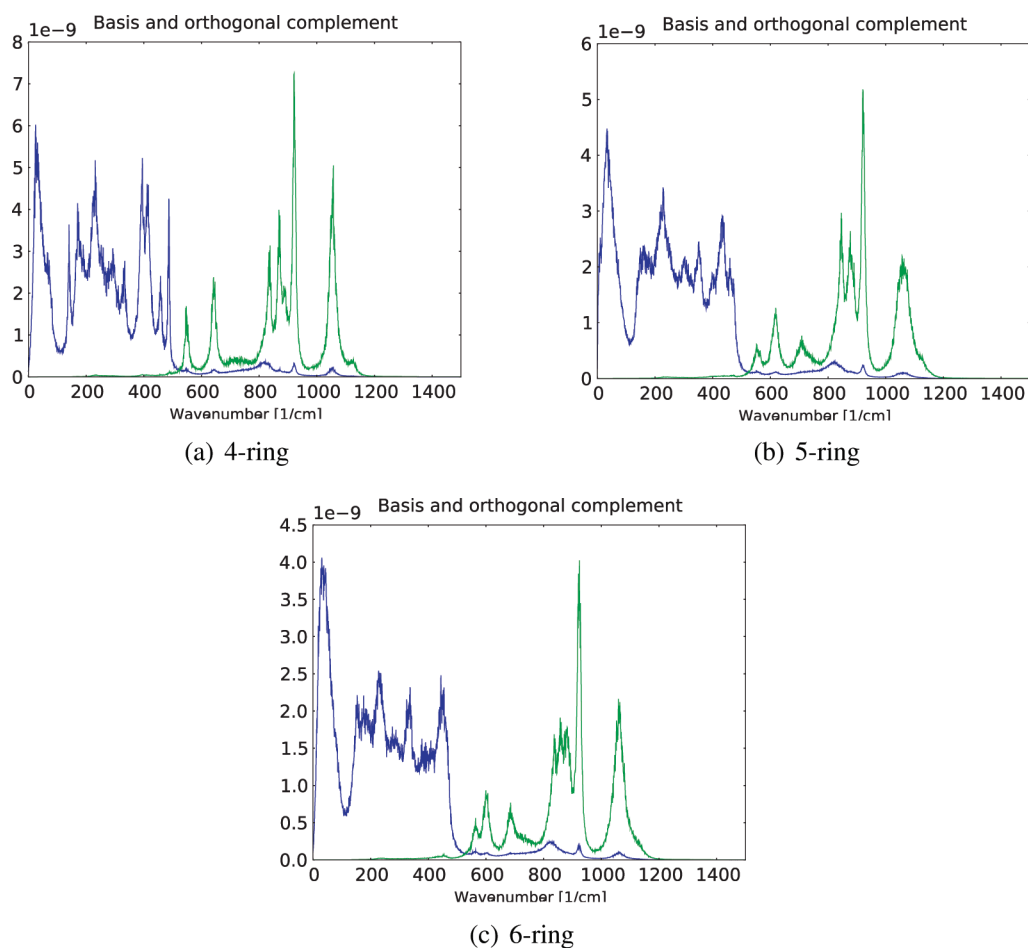


Figure 10. INS spectra of the projection of the atomic velocities on the basis of all O–Si IC's (green curve) and on the orthogonal complement of these IC's (blue curve).

the tangential vector of this puckering angle will reveal the real type of this specific mode.

Influence of the Connectivity. So far we studied the elementary four-, five-, and six-membered rings. We examined the influence of the sizes by comparing the spectra of them. Now we look at the connectivity (or the topology) by adding n -rings to these basic structures and investigate the corresponding fundamental changes in the spectra. In this section we will particularly focus on framework structures that contain five-membered rings (as in ref 26). Also the famous MFI-framework topology is built from five-membered rings. More precisely, in this section we will concentrate on the following 5T ring clusters: 7T(2×5); 8T(2×5); 8T(4×5); and 11T(3×5) (see Figure 1). They all show two or more connected five-rings except for the reference five-membered ring. Each mode ($q_{c,0}$, $\text{span}(q_{c,1}, q_{s,1})$, and $\text{span}(q_{c,2}, q_{s,2})$) is more or less degenerate due to the appearance of multiple five-rings, hence an average is taken of the corresponding spectra. Spectra belonging to other building blocks are given in the Supporting Information.

As already observed for the elementary four-, five-, and six-membered rings, the ring size will not affect the INS spectra (see Supporting Information) for the terminal O–Si and antisymmetric IC's. The changing topology induced by adding n -rings to elementary rings has no influence at all on the spectra; no peak shifts occur, and the shapes remain unaltered. This observation was more or less expected for the terminal O–Si stretches but

rather unexpected and even surprising for the antisymmetric stretch mode.

On the contrary, significant changes in the vibrational INS spectra occur for the various symmetric stretches. Spectra of the projected velocities on the circulant symmetric IC's of the five-membered ring systems under study are shown in Figure 11.

All spectra are referred to with respect to the reference spectral lines observed for the elementary five-membered ring. Two connected five-rings sharing two atoms (8T(2×5), Figure 1k) cause a splitting of the two main peaks with a slight blue and red shift for the $q_{c,0}$ and $\text{span}(q_{c,1}, q_{s,1})$ mode. In principle two degenerate $q_{c,0}$ modes should exist. The slight coupling between these modes (for two rings share two silica atoms) gives rise to the observed small splitting. A third peak is manifestly present in the breathing mode $q_{c,0}$ and could be resolved by a suitable selection of an additional IC. The other connected five-ring structures lead to similar features. The 11T(3×5) structure with three connected five-rings shows the same pattern as the 8T(2×5), with the difference that now three five-rings are attached to each other which is reflected in the projected spectrum. In the 7T(2×5) structure the situation becomes more complex because now three Si atoms are shared. Since in the 8T(4×5) structure even more five-rings are shared, the spectra become more distinct from the reference single five-ring spectra. For the $\text{span}(q_{c,2}, q_{s,2})$ mode, the spectra show a remarkable resemblance in all structures. Another issue concerns

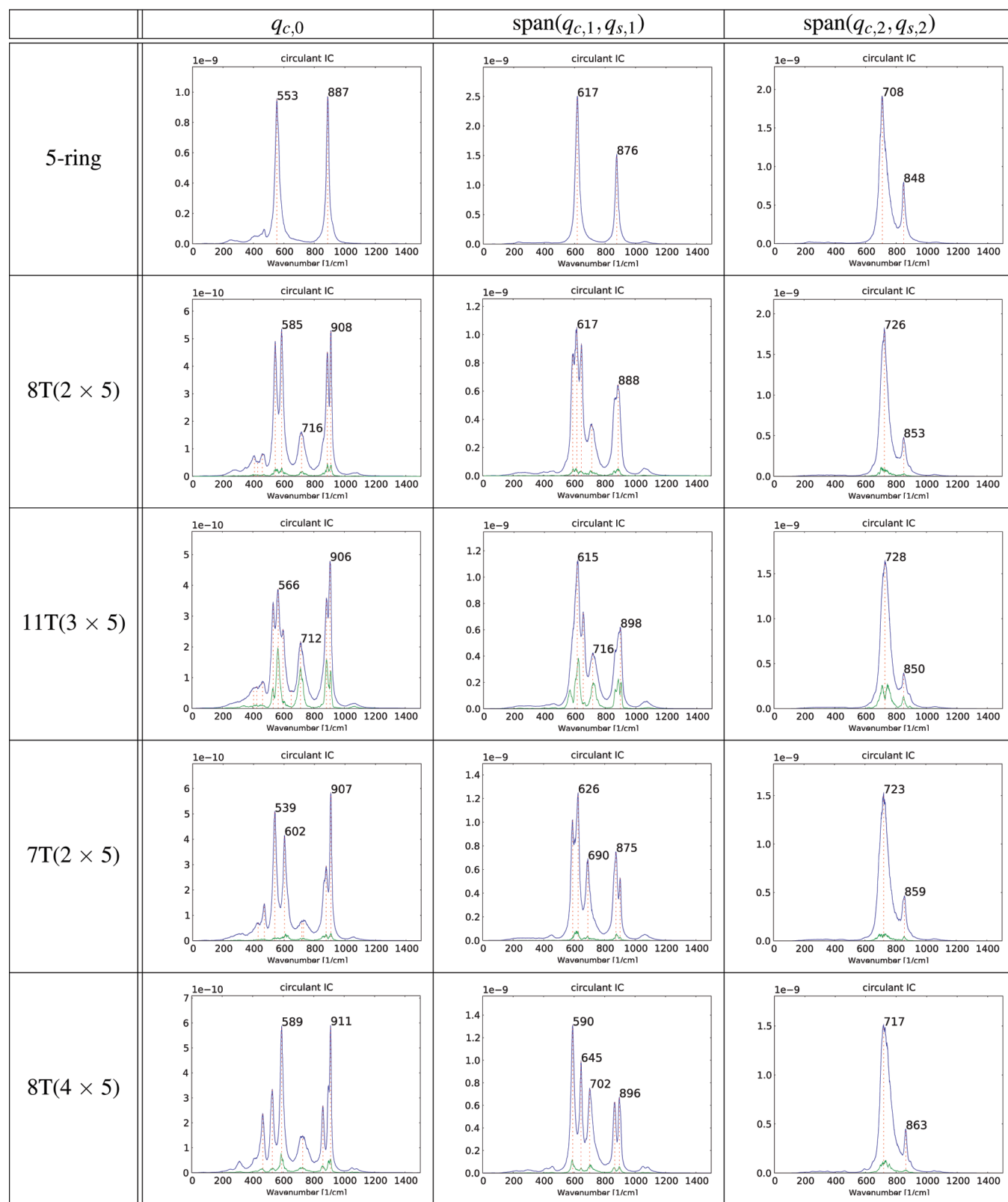


Figure 11. INS spectra belonging to atomic velocities projected on the circulant symmetric q_{ss} IC's for connected five-rings. Green curve: standard deviation.

the way spectra are affected by connecting rings of different sizes. We compare in Figure 12 the INS spectra generated by connecting a five-membered ring with, respectively, a four- and a six-

membered ring. Other spectra belonging to connected rings are shown in the Supporting Information. A common feature is the observation that when an n_2 -ring is attached to a basis n_1 -ring

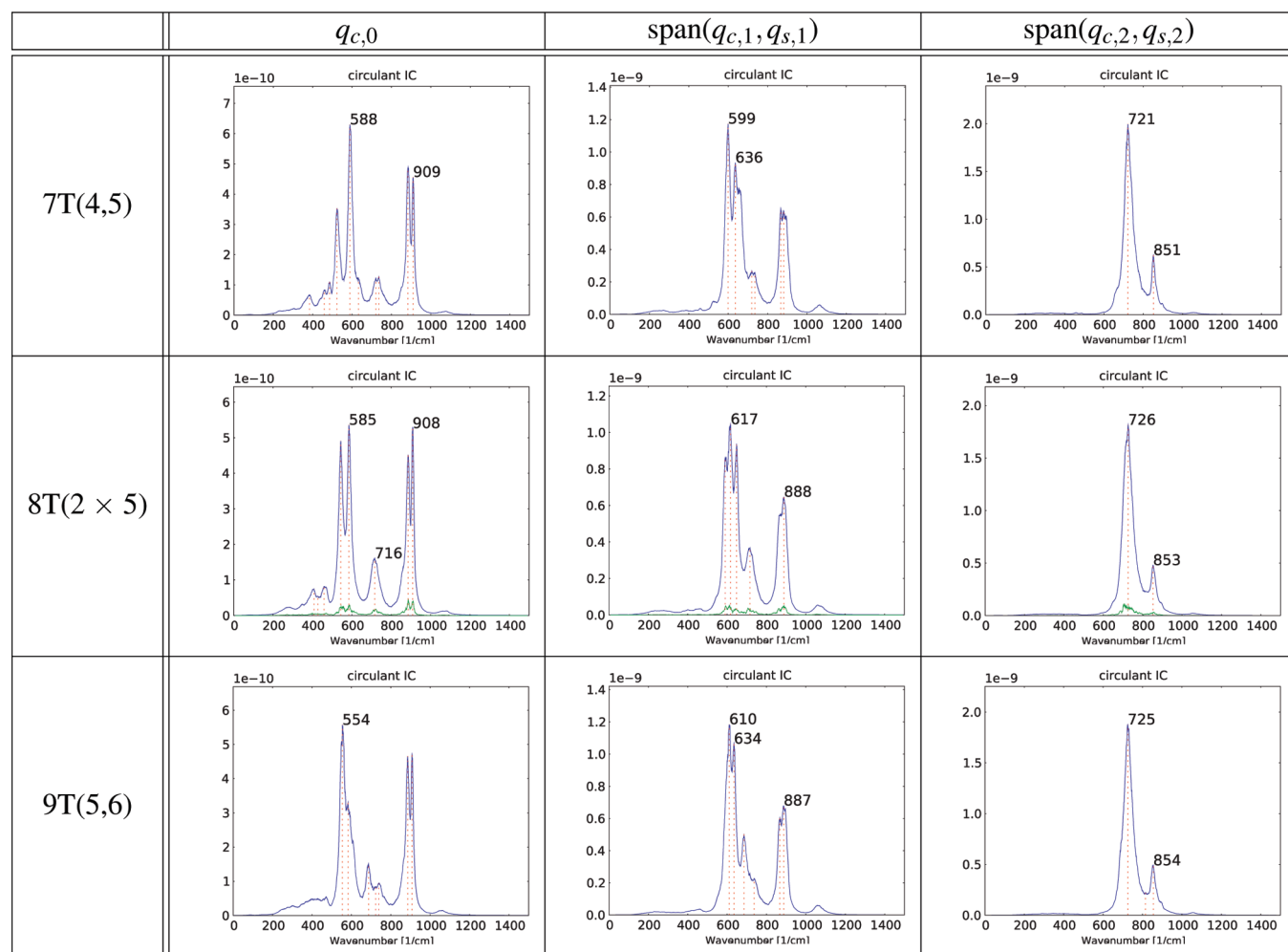


Figure 12. INS spectra belonging to atomic velocities projected on the circulant symmetric $q_{c,0}$, $\text{span}(q_{c,1}, q_{s,1})$ and $\text{span}(q_{c,2}, q_{s,2})$ IC's for the five-ring connected with, respectively, a four-, five-, and six-ring.

($n_1 = 4-6$), the peaks of the $q_{c,0}$ and $\text{span}(q_{c,1}, q_{s,1})$ mode of the basis n_1 -ring seem to split up. One can distinguish between symmetrical ($n_1 = n_2$, e.g., five-ring attached to five-ring) and antisymmetrical splitting ($n_1 \neq n_2$, e.g., four-ring attached to five-ring). These effects are a logical result from the fact that the peaks of the modes of the n_1 - and n_2 -rings occur at the same frequency. When the two rings are attached, a resonance effect arises which causes one peak to turn into two floating modes: one peak with a slightly higher and one peak with a slightly lower wavenumber around the resonance frequency. When $n_1 = n_2$ the peaks show more or less the same amplitude, when $n_1 \neq n_2$ the difference in amplitude is much larger. Therefore it is useful to introduce the nomenclature of symmetrical and antisymmetrical splitting. An example is the $7T(2 \times 5)$: the 553 and 887 cm^{-1} peaks of the five-ring split into 539, 602, 876, and 907 cm^{-1} , respectively. Just as in the case of connecting five-rings, the $\text{span}(q_{c,2}, q_{s,2})$ mode remains the same.

The $q_{c,0}$ breathing mode is the most sensitive to connectivity differences. When more rings are attached to the elementary n_1 -rings, it is obvious that more peaks appear, since the ring systems cannot be regarded as independent systems anymore and mixing of modes occurs which results in extra peaks. This coupling gives rise to a splitting of the original peaks or the emerging of additional peaks. The $q_{c,0}$ mode is the most collective

mode because the breathing mode is determined by stretches which are not in an immediate antiphase (see Figure 5), giving this mode a more global character. The eigenmode of the $\text{span}(q_{c,2}, q_{s,2})$ IC on the other hand is the most local mode, as the successive stretches move in antiphase with each other (Figure 5). The topology dependence here is minimal.

It is important to note that the $q_{c,0}$ breathing mode is not affected by the size of the ring (see subsection "Circulant Symmetric IC's"), but is affected by the particular way in which it is connected to other rings.

CONCLUSION AND PERSPECTIVES

In this paper we have used the velocity projection method to assign (parts of) spectral peaks in zeolite vibrational spectra to particular changes in IC's. The focus was on essential zeolite building units, which can be assumed to play important roles during initial states of zeolite formation and during zeolite growth in general. The analysis of vibrational spectra can be a very useful tool in understanding the process, as the spectral behavior of eigenmodes of IC's can vary with different molecular systems. MD can then be used to simulate various structures which are key components during zeolite growth. The atomic velocities obtained from these MD runs can then be projected on a well chosen set of IC's.

Here, we focused on stretches, linear combinations of them, and eigenvectors of the circulant matrix as IC's. The circulant matrix was proposed as a model for the Hessian of ring molecular structures.

We found that the spectra of the terminal O–Si and the antisymmetric internal O–Si IC's were not influenced by the topology and by the size of the considered rings; the spectra of the elementary four-, five-, and six-ring are the same. In addition, when other rings are attached to these basic n -rings, the same conclusions can be drawn.

When the atomic velocities are projected on the symmetric O–Si IC's, the resulting spectra differ for the four-, five-, and six-ring. The spectral shifts of the spectra of IC $q_{c,0}$ are tiny, while those corresponding with $\text{span}(q_{c,m}, q_{s,m})$ are rather substantial. The same trend can be observed in connected ring systems. We should, however, be very careful in drawing conclusions, as it is unclear what the role is of the force field in this specific field. It is recommended to refer to the spectra of the elementary rings when investigating the effect of connecting multiple n -rings. The most important conclusions are that the $q_{c,0}$ IC is the most sensitive mode to changes in molecular topology. It is the most global mode. We associate it to the breathing mode as it is determined by adjacent stretches which are not in antiphase. All the peaks occurring in the projected velocity $q_{c,0}$ power spectra are retrieved in power spectra belonging to other internal coordinates defining the ring-opening vibration.

About the connectivity we found that the size of the added n -rings does not play a crucial role; it is rather the way in which these rings are connected to each other that is the origin of the observed spectral changes.

When one projects the atomic velocities on the orthogonal complement of all stretch IC's present in a molecular structure, we found that the contribution of the stretch modes to the spectra is spectrally well separated from other contributions, like bending and dihedral angles. This confirms that the spectral region in which we are interested for zeolite synthesis can be almost completely resolved by considering changes in stretch modes.

In an attempt to decompose the spectra of the symmetric IC's further we introduced the circulant matrix as a model for the Hessian of the considered ring systems. If we want to extend the applications to systems which are not restricted to a ring structure, another more general approach is desired. A first incentive for this has already been found by linking the displacement (small motions) in IC's with the displacement (small motions) in Cartesian coordinates. We represent the displacement of the IC's by the $3N \times K$ matrix J_{jk} (N is the number of atoms, K is number of IC's) and the Cartesian displacement by the $3N \times 3N$ matrix $E_{i\lambda}$. The Cartesian eigenmodes can then be expanded in a redundant set of internal coordinates (they are not orthogonal) with $\alpha_{k\lambda}$ being the matrix which holds the expansion coefficients:

$$\sum_{k=1}^K J_{jk} \alpha_{k\lambda} = E_{i\lambda} \quad (27)$$

Here λ is a counter for the number of eigenmodes. The coefficients of the matrix $\alpha_{k\lambda}$ can then be used as coefficients for projecting the atomic velocities on the considered internal coordinates. The Cartesian eigenmodes $E_{i\lambda}$ can only be obtained if the Hessian of the potential energy function is diagonalized, thus the mass-weighted normal mode eigenvalue equation has to be solved. To obtain the Hessian, a specified energy function V has to be provided. First results of this method were obtained, but it is still a work in progress.

APPENDIX

Eigenvalues and the k^{th} component of the normalized eigenvectors. We will show that component k of eigenvector m has the following form:

$$V_k^{(m)} = \exp\left(\frac{2\pi imk}{n}\right) \quad (28)$$

It is easy to see that the set of n orthogonal vectors $V^{(m)}$ with components of the form of eq 28, are eigenvectors of C . The product of C with $V^{(m)}$ can be written as

$$\begin{aligned} \sum_k C_{jk} V_k^{(m)} &= \sum_k c_{(j-k)} \exp\left(\frac{2\pi imk}{n}\right) \\ &= \sum_{k'} c_{(k')} \exp\left(\frac{2\pi im(j-k')}{n}\right) \\ &= \left[\sum_{k'} c_{(k')} \exp\left(\frac{-2\pi imk'}{n}\right) \right] \exp\left(\frac{2\pi imj}{n}\right) \\ &= \lambda^{(m)} V_j^{(m)} \end{aligned} \quad (29)$$

where the dummy summation index k was changed to $k' = (j - k) \bmod n$. It is clear that $V^{(m)}$ is indeed eigenvectors of the symmetric circulant matrix C .

The eigenvalues are real and can straightforwardly be rewritten as

$$\begin{aligned} \lambda^{(m)} &= \sum_k c_{(k)} \exp\left(\frac{-2\pi imk}{n}\right) \\ &= \sum_k \frac{1}{2} (c_{(k)} + c_{(-k)}) \exp\left(\frac{-2\pi imk}{n}\right) \\ &= \sum_k c_{(k)} \frac{1}{2} \left[\exp\left(\frac{-2\pi imk}{n}\right) + \exp\left(\frac{2\pi imk}{n}\right) \right] \\ &= \sum_k c_{(k)} \cos\left(\frac{2\pi mk}{n}\right) = \lambda^{(-m)} \end{aligned} \quad (30)$$

ASSOCIATED CONTENT

S Supporting Information. Additional projected velocity power spectra, IR and INS spectra of all the oligomers are given. This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

E-mail: veronique.vanspeybroeck@ugent.be.

ACKNOWLEDGMENT

This work is supported by the Fund for Scientific Research—Flanders (FWO), the Research Board of Ghent University (BOF) and BELSPO in the frame of IAP/6/27, the Belgian Prodex office, and ESA. Part of the computational resources and services used in this work were provided by Ghent University. V.

V.S. acknowledges the European Research Council under the European Community's Seventh Framework Programme (FP7- (2007-2013) ERC grant agreement no. 240483). J.A.M. acknowledges the Flemish government for long term structural funding (Methusalem).

REFERENCES

- (1) Boronat, M.; Viruela, P.; Corma, A. *J. Am. Chem. Soc.* **2003**, *126*, 3300–3309.
- (2) Nascimento, M. *J. Mol. Struct.* **1999**, *464*, 239–247.
- (3) Demontis, P.; Suffritti, G. *Chem. Rev.* **1997**, *97*, 2845–2878.
- (4) Hedlund, J.; Schoeman, B.; Sterte, J. *Progress in Zeolites and Microporous Materials*; Chon, H., Ihm, S.-K., Uh, Y. S., Eds.; Elsevier: Amsterdam, The Netherlands, 1997; pp 2203–2210.
- (5) Ravishankar, R.; Kirschhock, C.; Schoeman, B.; Vanoppen, P.; Grobet, P.; Storck, S.; Maier, W.; Schryver, F. D.; Martens, J.; Jacobs, P. *J. Phys. Chem. B* **1998**, *102*, 2633–2639.
- (6) Dokter, W.; van Garderen, H.; Beelen, T.; van Santen, R.; Bras, W. *Angew. Chem., Int. Ed.* **1995**, *34*, 73–75.
- (7) Knight, C.; Wang, J.; Kinrade, S. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3099–3103.
- (8) Auerbach, S.; Monson, P.; Ford, M. *Curr. Opin. Colloid Interface Sci.* **2005**, *10*, 220–225.
- (9) Lewis, D.; Richard, C.; Catlow, A.; Thomas, J. *Faraday Discuss.* **1997**, *106*, 451–471.
- (10) Burkett, S.; Davis, M. *Chem. Mater.* **1995**, *7*, 920–928.
- (11) Corkery, R.; Ninham, B. *Zeolites* **1997**, *18*, 379–386.
- (12) Schoeman, B. *Microporous and Mesoporous Mater.* **1998**, *22*, 9–22.
- (13) Tsay, C.; Chiang, A. *Microporous Mesoporous Mater.* **1998**, *26*, 89–99.
- (14) Watson, J.; Brown, A.; Iton, L.; White, J. *J. Chem. Soc., Far. Trans.* **1998**, *94*, 2181–2186.
- (15) de Moor, P.; Beelen, T.; van Santen, R.; Beck, L.; Davis, M. *J. Phys. Chem. B* **2000**, *104*, 7600–7611.
- (16) Li, Q.; Mihailova, B.; Creaser, D.; Sterte, J. *Microporous and Mesoporous Mater.* **2001**, *43*, 51–59.
- (17) Mintova, S.; Olson, N.; Senker, J.; Bein, T. *Angew. Chem., Int. Ed.* **2002**, *41*, 2558–2561.
- (18) Aerts, A.; Follens, L.; Haouas, M.; Caremans, T.; Delsuc, M.; Loppinet, B.; Vermant, J.; Goderis, B.; Taulelle, F.; Martens, J.; Kirschhock, C. *Chem. Mater.* **2007**, *19*, 3448–3454.
- (19) Kumar, S.; Davis, T.; Ramanan, H.; Penn, R.; Tsapatsis, M. *J. Phys. Chem. B* **2007**, *111*, 3398–3403.
- (20) Jin, L.; Auerbach, S.; Monson, P. *J. Phys. Chem. C* **2010**, *114*, 14393–14401.
- (21) Kirschhock, C.; Ravishankar, R.; Looveren, L. V.; Jacobs, P.; Martens, J. *J. Phys. Chem. B* **1999**, *103*, 4972–4978.
- (22) Kirschhock, C.; Ravishankar, R.; Verspeurt, F.; Grobets, P.; Jacobs, P.; Martens, J. *J. Phys. Chem. B* **1999**, *103*, 4965–4978.
- (23) Ravishankar, R.; Kirschhock, C.; Knops-Gerrits, P.; Feijen, E.; Grobet, P.; Vanoppen, P.; Schryver, F. D.; Mieke, G.; Fuess, H.; Schoeman, B.; Jacobs, P.; Martens, J. *J. Phys. Chem. B* **1999**, *103*, 4960–4964.
- (24) Verstraelen, T.; Szyja, B.; Lesthaeghe, D.; Declerck, R.; Van Speybroeck, V.; Waroquier, M.; Jansen, A.; Aerts, A.; Follens, L.; Martens, J.; Kirschhock, C.; van Santen, R. *Top. Catal.* **2009**, *52*, 1261–1271.
- (25) Petry, D.; Haouas, M.; Wong, S.; Aerts, A.; Kirschhock, C.; Martens, J.; Gaskell, S.; Anderson, M.; Taulelle, F. *J. Phys. Chem. C* **2009**, *113*, 20827–20836.
- (26) Lesthaeghe, D.; Vansteenkiste, P.; Verstraelen, T.; Ghysels, A.; Kirschhock, C.; Martens, J.; Van Speybroeck, V.; Waroquier, M. *J. Phys. Chem. C* **2008**, *112*, 9186.
- (27) In *Normal mode analysis: theory and applications to biological and chemical systems*; Cui, Q., Bahar, I., Ed.; Chapman & Hall: Boca Raton, FL, 2006.
- (28) Scribano, Y.; Benoit, D. *M. J. Chem. Phys.* **2007**, *127*, 164118.
- (29) Bornhauser, P.; Calzaferri, G. *J. Phys. Chem.* **1996**, *100*, 2035–2044.
- (30) Pasquarello, A.; Car, R. *Phys. Rev. Lett.* **1998**, *80*, 5145–5147.
- (31) Smirnov, K.; Bougeard, D. *Catal. Today* **2001**, *70*, 243–253.
- (32) To, T.; Bougeard, D.; Smirnov, K. *J. Raman Spectrosc.* **2008**, *39*, 1869–1877.
- (33) Arab, M.; Bougeard, D.; Smirnov, K. *Phys. Chem. Chem. Phys.* **2002**, *4*, 1957–1963.
- (34) Jobic, H.; Smirnov, K.; Bougeard, D. *Chem. Phys. Lett.* **2001**, *344*, 147–153.
- (35) Smirnov, K.; Bougeard, D. *J. Raman Spectrosc.* **1993**, *24*, 255–257.
- (36) Smirnov, K.; Bougeard, D. *J. Phys. Chem.* **1993**, *97*, 9434–9440.
- (37) Smirnov, K.; Bougeard, D.; Maire, M. L.; Brémard, C. *Chem. Phys.* **1994**, *179*, 445–454.
- (38) Kuhne, T.; Krack, M.; Mohamed, F.; Parrinello, M. *Phys. Rev. Lett.* **2007**, *98*, 066401.
- (39) Kuhne, T.; Krack, M.; Parrinello, M. *J. Chem. Theory Comput.* **2009**, *5*, 235–241.
- (40) Jacob, C.; Reiher, M. *J. Chem. Phys.* **2009**, *130*, 084106.
- (41) Verstraelen, T.; Neck, D. V.; Ayers, P.; Speybroeck, V. V.; Waroquier, M. *J. Chem. Theory Comput.* **2007**, *3*, 1420.
- (42) Huang, Y.; Jiang, Z. *Microporous Mater.* **1997**, *12*, 341–345.
- (43) Flanigen, E.; Khatami, H.; Szymanski, H. In *Infrared Structural Studies of Zeolite Frameworks*, Advances in Chemistry Series; Flanigen, E. M., Sand, L. B., Eds.; American Chemical Society: Washington, D.C., 1974; Vol. 101.
- (44) de Man, A.; van Santen, R. *Zeolites* **1992**, *12*, 269–278.
- (45) van Santen, R.; Vogel, D. *Adv. Solid-State Chem.* **1989**, *1*, 151–224.
- (46) Wood, S. *J. R. Stat. Soc., B* **2000**, *62*, 413.
- (47) CP2K, 2008; <http://cp2k.berlios.de>.
- (48) Verstraelen, T.; Van Speybroeck, V.; Waroquier, M. *J. Chem. Inf. Mod.* **2008**, *48*, 1530–1541.
- (49) Verstraelen, T.; Van Houteghem, M.; Van Speybroeck, V.; Waroquier, M. *J. Chem. Inf. Mod.* **2008**, *48*, 2414.
- (50) Futrelle, R.; McGinty, D. *Chem. Phys. Lett.* **1971**, *12*, 285–287.
- (51) McQuarrie, D. In *Statistical mechanics*; Cato, R., Ed.; Harper & Row, 1976.
- (52) Berens, P.; Wilson, K. *J. Chem. Phys.* **1981**, *74*, 4872–4882.
- (53) Noid, D.; Koszykowski, M.; Marcus, R. *J. Chem. Phys.* **1977**, *67*, 404.
- (54) Papoulis, A. In *Probability, random variables and stochastic processes*; York, N., Ed.; McGraw-Hill, 1965.
- (55) Urey, H.; Bradley, C. *Phys. Rev.* **1931**, *38*, 1969–1978.
- (56) Watson, J.; Iton, L.; Keir, R.; Thomas, J.; Dowling, T.; White, J. *J. Phys. Chem. B* **1997**, *101*, 10094–10104.
- (57) Kragten, D.; Fedeyko, J.; Sawant, K.; Rimer, J.; Vlachos, D.; Lobo, R.; Tsapatsis, M. *J. Phys. Chem. B* **2003**, *107*, 10006–10016.
- (58) Mohamed, R.; Aly, H.; El-Shahat, M.; Ibrahim, I. *Microporous and Mesoporous Mater.* **2005**, *79*, 7–12.
- (59) Haouas, M.; Taulelle, F. *J. Phys. Chem. B* **2006**, *110*, 3007–3014.
- (60) Moravetski, V.; Hill, J.-R.; Eichler, U.; Cheetham, A.; Sauer, J. *J. Am. Chem. Soc.* **1996**, *118*, 13015–13020.
- (61) Press, W.; Teukolsky, S.; Vetterling, W.; Flannery, B. In *Numerical recipes in C: The art of scientific computing*; Press, W., Ed.; Cambridge University Press: New York, 1992.

Spectroscopic Properties of Azobenzene-Based pH Indicator Dyes: A Quantum Chemical and Experimental Study

Daniel Escudero,[†] Sabine Trupp,[‡] Beate Bussemer,[§] Gerhard J. Mohr,[‡] and Leticia González^{*,†}

[†]Institute of Physical Chemistry, Theoretical Chemistry, Friedrich-Schiller University Jena, Helmholtzweg 4, D-07743 Jena, Germany

[‡]Department of Polytronic Systems, Workgroup Sensor Materials, Fraunhofer Research Institution for Modular Solid State Technologies (EMFT), Josef-Engert Strasse 9, D-93053 Regensburg, Germany

[§]Institute of Physical Chemistry, Friedrich-Schiller University Jena, Lessingstrasse 10, D-07743 Jena, Germany

ABSTRACT: The UV–visible absorption spectra of six new optical sensors based on acidochromic azobenzenes have been measured and assigned with the help of quantum chemical calculations. The investigated compounds are able to monitor the pH in the range from pH 3–10. Using the hybrid density functional PBE0 and including solvent effects with a polarized continuum model, the agreement between the experimental and theoretical UV/vis spectra of the dyes in their neutral and anionic forms is very good. The spectroscopic $\pi\pi^*$ states, responsible for the optical properties of the sensors, are described within an accuracy of 0.1 eV. Similar accuracy is demonstrated in the $n\pi^*$ states. The $\pi\pi^*$ states can be assigned as a charge transfer from the aromatic π orbital localized in the azo-phenol moiety to the antibonding π^* of the azo group. Under basic conditions, the spectrum is bathochromically shifted and more intense than in acid media. Upon substitution in the phenyl moiety, red- or blue-shifts of the UV–visible bands are observed depending on whether the substituent is electron-donor or -withdrawing, respectively. These effects are stronger at high pH values and can be rationalized in terms of the stabilization and/or destabilization of the involved frontier orbitals.

1. INTRODUCTION

The development of pH sensors is a continuous challenge in chemistry.¹ Although the determination of pH with traditional electrochemical sensors is well-established, optical sensors are a valuable alternative where glass electrodes are impractical or impossible to use.² Moreover, optical sensors can be more versatile than electrodes, as they are easy and inexpensive to fabricate,³ and profit from the current advances of optical spectroscopy.

Azobenzene and derivatives have attracted a considerable amount of attention due to their capability to reversibly switch between the cis and trans conformers using two different wavelengths,⁴ and therefore their large applicability as molecular devices.⁵ They are used as light-driven membranes,⁶ as single-molecule optomechanical machines,⁷ as media storage,⁸ or to control peptide folding.⁹ Moreover, because they allow for facile and multiple functionalizations, a wide range of azobenzene-based indicator dyes have been synthesized, initially for detecting alkali, earth alkali, and heavy metal ions,¹⁰ but later also for monitoring electrically neutral and anionic analytes such as alcohols,¹¹ amines,¹² aldehydes,¹³ saccharides,¹⁴ and bisulfite.¹⁵ Recently, some of us have synthesized new derivatives of 2-hydroxyethylsulfonyl azobenzene (HESAB) indicator dyes with emphasis on measuring the pH in range between 3 and 10.¹⁶ These indicators can be covalently linked to polymers containing hydroxyl functions such as cellulose, polyurethane hydrogel, and hydroxyalkyl methacrylate. Because one of our aims is to design indicator dyes, which exhibit strong color changes in this pH range, the goal of this Article is to present a comprehensive study of the substitution effects on the spectral properties of the corresponding protonated and deprotonated forms of HESAB. In particular, we want to predict whether or not

the absorbance spectra of the phenolic dyes in going from acid to base form (i.e., protonated to deprotonated form) are well separated, thus providing indicator materials whose color changes can be easily distinguished and quantified by optical sensor modules.

To shed some light into the photophysics of the HESAB compounds, which can in turn help to predict the potential use of these dyes as pH indicators, the experimental spectra have been characterized with the help of quantum chemical calculations. While recording absorption spectra can be close to routine, the computation of electronically excited states of large organic dyes with chemical accuracy (less than 0.1 eV) is still a challenging problem in modern quantum chemistry. Currently, time-dependent density functional theory (TD-DFT) is the most extended theoretical tool to compute transition energies and oscillator strengths in organic and inorganic compounds.¹⁷ Although the performance of different functionals may vary depending on the treated systems,¹⁸ it seems that corrected hybrid functionals are best suited to describe the excited states of delocalized aromatic dye systems,¹⁹ and this approach is used here.

2. EXPERIMENTAL SECTION

Sample Preparation. The six azobenzene dyes (1–6) with their deprotonated forms (1[−]–6[−]) shown in Figure 1 were investigated. The azo function is in para position to the phenol hydroxyl group. Substitution effects are studied by substituting H by two electron donors (CH₃, OCH₃) and/or two electron-withdrawing (F, Br) substituents on the phenyl moiety.

Received: December 16, 2010

Published: March 09, 2011

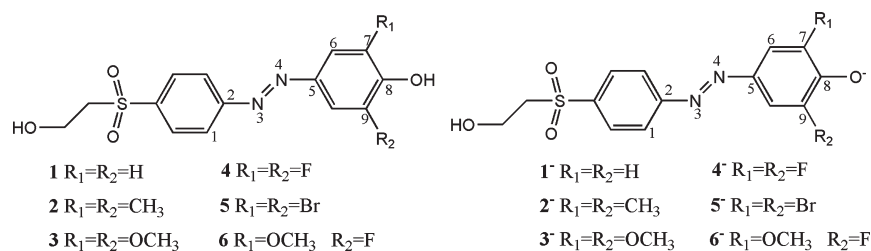


Figure 1. Chemical structure of the 2-hydroxyethylsulfonyl azobenzene dyes here investigated.

Table 1. Main Geometrical Parameters of Gas-Phase-Optimized Compounds 1–6 and 1⁻–6⁻^a

compound	d_{3-4}^b	d_{2-3}^b	d_{4-5}^b	d_{6-7}^b	d_{8-9}^b	α_{2-3-4}^c	α_{3-4-5}^c	$\tau_{1-3-4-6}^c$
azobenzene								
RI-BP86/TZVP ^d	1.267	1.417	1.417			114.8	114.8	180
PBE0/TZVP	1.243	1.413	1.413	1.386	1.393	115.1	115.1	180
X-ray ^e	1.247	1.428	1.428	1.384	1.391	114.1	114.1	178.1
1 (PCM-PBE0/TZVP)	1.248	1.413	1.399	1.380	1.401	114.7	116.2	170.5
1 (PBE0/TZVP)	1.271	1.418	1.408	1.390	1.410	114.2	115.4	167.5
2	1.246	1.412	1.402	1.384	1.405	114.5	115.7	163.2
3	1.247	1.410	1.400	1.388	1.404	114.3	115.8	162.3
4	1.244	1.412	1.405	1.377	1.397	114.6	115.2	165.7
5	1.244	1.412	1.405	1.377	1.397	114.6	115.2	165.7
6	1.245	1.412	1.403	1.386	1.392	114.4	115.6	166.2
1 ⁻ (PCM-PBE0/TZVP)	1.276	1.395	1.361	1.368	1.448	114.0	117.2	178.8
1 ⁻ (PBE0/TZVP)	1.288	1.377	1.344	1.359	1.459	113.7	117.4	173.1
2 ⁻ (PCM-PBE0/TZVP)	1.282	1.390	1.357	1.369	1.457	114.0	117.2	178.6
2 ⁻ (PBE0/TZVP)	1.289	1.376	1.344	1.361	1.465	113.7	117.5	172.9
3 ⁻	1.295	1.370	1.339	1.361	1.471	113.8	117.1	172.6
4 ⁻	1.283	1.380	1.347	1.355	1.459	113.8	117.1	171.4
5 ⁻	1.277	1.386	1.354	1.359	1.466	113.9	117.2	170.9
6 ⁻	1.289	1.375	1.343	1.363	1.456	113.9	117.2	171.8

^aTheoretical and experimental values of *trans*-azobenzene are given for comparison. The corresponding chemical structures with the atom numbering can be found in Figure 1. ^bDistances in angstroms. ^cAngles in degrees. ^dValues obtained under C_{2h} symmetry constraint from ref 29. ^eExperimental values from ref 28.

The synthesis of the 2-hydroxyethylsulfonyl azo dye **1** was accomplished by diazotation of 2-(4-aminobenzenesulfonyl)ethanol to phenol in acetic acid. The product was precipitated as fine yellow-orange powder. Yield: 28%. Anal. Calcd for C₁₄H₁₄N₂O₄S (306.34 g/mol): C, 54.89%; H, 4.61%; N, 9.14%; S, 10.47%. Found: C, 54.89%; H, 4.53%; N, 9.10%; S, 10.72%. ¹H NMR (DMSO): δ (ppm) 7.85–8.07 (m, 6H, =CH–), 6.98 (m, 2H, =CH–), 3.74 (t, 2H, –CH₂–), 3.51 (t, 2H, –CH₂–). Mass spectrometry data (EI): m/e (%), 306 (38). The synthesis of the 2-hydroxyethylsulfonyl azo dyes **2–6** has been described in detail elsewhere.¹³

Spectroscopic Measurements. The absorbance spectra of the dissolved dyes **1–6** were recorded on a Lambda 16 UV–vis spectrometer (Perkin-Elmer) at 20 ± 2 °C. The dyes were dissolved in methanol and mixed with the aqueous buffered solutions in a 1:1 ratio because of the low solubility of the dyes in pure aqueous solution. A wide pH range buffer was used, which was 0.04 M in sodium acetate, 0.04 M in boric acid, 0.04 M in sodium dihydrogen phosphate, and 0.1 M in sodium sulfate. The pH was adjusted in a way that exclusively the neutral or the anionic form of the dyes was observed.

3. COMPUTATIONAL DETAILS

For the sake of computational ease, the 2-hydroxyethyl function was replaced by a methyl group in all of the calculations (see Figure 1). The resulting geometries were optimized in the *trans* configuration using the global hybrid functional PBE0²⁰ in combination with a polarized valence triple- ζ basis set (TZVP) for all atoms. The nature of the stationary points was confirmed by calculating the Hessian at the same level of theory. The calculation of the transition energies and oscillator strengths was done with the same functional. With 25% of exact Hartree–Fock exchange, PBE0 was chosen because it delivers excitation energies with mean absolute errors of 0.14 eV for organic dyes.^{19b} The effect of the methanol environment on the excitations energies was modeled with the polarized continuum model (PCM)²¹ and $\epsilon = 32.6$. For reference, the UV spectrum of the unsubstituted compounds **1/1⁻** is also calculated using correlated *ab initio* second-order approximated coupled-cluster theory²² in the resolution of the identity approximation (RI-CC2)²³ and the pure functional BP86^{24,25} also within RI approximation. The effect of the solvent on the geometries was also investigated at the PBE0 level of theory. RI-BP86 and

Table 2. UV/Vis Experimental Data of HESAB Compounds 1–6 under Acid and Basic Conditions (pH Is Also Given) Measured in Buffered Methanol (v/v = 1:1)^a

	1 λ_{\max}	2 λ_{\max}	3 λ_{\max}	4 λ_{\max}	5 λ_{\max}	6 λ_{\max}
acid	pH = 3.52 456, 361 (2.72, 3.44)	pH = 3.17 373 sh, 427 (3.32 sh, 2.91)	pH = 4.16 460 sh, 395 (2.70 sh, 3.14)	pH = 3.19 355 (3.49)	pH = 3.19 354 (3.50)	pH = 3.16 374 (3.32)
basic	pH = 11.08 460 (2.70)	pH = 11.28 499 (2.49)	pH = 10.83 525 (2.36)	pH = 11.44 444 (2.79)	pH = 11.54 460, 292 (2.70, 4.25)	pH = 11.28 487(2.55)

^a Maximum absorption peaks and most relevant shoulders (sh) in nm and in eV in parentheses.

Table 3. Molar Extinction Coefficients of the Dyes in Pure Methanol^a

	ϵ [L mol ⁻¹ cm ⁻¹]		λ_{\max} neutral form [nm]	λ_{\max} anionic form [nm]
	neutral form	anionic form		
1/1 ⁻	27 300	27 700	364	441
2/2 ⁻	24 190	29 240	373	498
3/3 ⁻	18 890	36 090	396	522
4/4 ⁻	25 200	30 000	354	440
5/5 ⁻	14 430	21 740	354	452
6/6 ⁻	19 740	30 730	375	485

^a The neutral form was obtained by addition of 100 μ L of 0.1 M hydrochloric acid, while the anionic form was obtained by addition of 100 μ L of triethylamine.

RI-CC2 calculations were performed with the TURBOMOLE.5.10 program package,²⁶ while the rest of calculations were performed with Gaussian 03.²⁷

4. RESULTS AND DISCUSSION

Structure of Azo-vinyl Sulfonyl Dyes. The most relevant geometrical parameters of the studied compounds 1–6 and corresponding anions are collected in Table 1. For the sake of comparison, azobenzene in the *trans*-conformation is also included. The low-lying excited states of azobenzene (and derivatives) involve the antibonding orbital located in the N=N bond; therefore, it is important to describe this bond length properly. Experimentally, the N=N distance of azobenzene is 1.247 Å.²⁸ Hättig et al. showed that triple- ζ basis sets are indispensable for an accurate description of this bond.²⁹ It has been estimated that RI-BP86/TZVP level of theory delivers bond distances in azobenzene ca. 0.001 Å larger than MP2 but very similar to the crystal structure.²⁹ As we see in Table 1, RI-BP86/TZVP overestimates the N=N bond (d_{3-4}) by ca. 0.02 Å. Interestingly, the result of the hybrid functional PBE0 (1.243 Å) is in much better agreement with the X-ray distance. The adjacent N–C bonds (d_{2-3} and d_{4-5}) are underestimated by both BP86 and PBE0 by ca. 0.01 Å. Despite being noticeable, these discrepancies are within the error bar of the experiment.²⁸

In view of the previous results and because no X-ray structures are available for HESAB derivatives to compare, we shall consider 0.01 Å as the error bar for the geometries of the HESAB compounds obtained with PBE0/TZVP. The N=N bond in the unsubstituted HESAB compound 1 is larger than in azobenzene and further destabilized upon deprotonation. Interestingly, when the solvent is included in the optimization (PCM-PBE0 values), the azo bond (d_{3-4}) is compressed in the neutral and anionic forms, while the bonds nearby (d_{2-3} and d_{4-5}) are

almost unaffected in the neutral form but stretched by ca. 0.02 Å in the anion.

The changes in the bond angles are negligible in both neutral and ionic forms (see α values in Table 1). In contrast, the planarity of these compounds deserves some attention. Despite some controversy,³⁰ nowadays it seems well-accepted that azobenzene is planar in the gas phase, as suggested by early X-ray experiments²⁸ and confirmed by accurate MP2 calculations.^{29,31} Simple substituted 4,4'-azobenzene derivatives are also planar at both MP2 and DFT levels of theory if the basis set is sufficiently large. On the contrary, if a smaller basis set, that is, MP2/6-31+G(d) level of theory is used, significantly twisted geometries, with a dihedral angle of ca. 20°, are predicted.³⁰ The dyes investigated here are twisted by ca. 15° (see dihedral $\tau_{1-3-4-6}$ in Table 1). Because we employ the flexible TZ basis set, we are confident that the nonplanarity of HESAB systems in vacuum is physical and due to the bulky nature of the substituents. We note that the frequencies corresponding to the out-of-plane motion of the rings are very small (e.g., 62 and 105 cm⁻¹ in compound 1 and of similar magnitude or even smaller in the other neutral species), increasing the possibility that nonplanar geometries contribute to the experimental absorption spectrum (see below). Worth mentioning is that a certain degree of planarity is recovered upon deprotonation and that the 1/1⁻ and 2⁻ optimized structures in the presence of solvent are slightly more planar than in the gas phase (see Table 1). These effects are especially important to address properly the $n\pi^*$ excitations of HESAB compounds, as we shall discuss below.

The rest of the compounds have been optimized only at the PBE0/TZVP level of theory. In the neutral HESAB, the calculated values for N=N and adjacent C–N bonds (now asymmetric), as well as the angle between them, are very similar to those of azobenzene (see Table 1). Both the electron donor (2 and 3) and the electron acceptor (4 and 5) substituents have little effect on the geometry of the neutral (or the anionic) HESAB. However, there are interesting changes upon deprotonation. In the phenolate form, a strong resonance effect through the conjugated system can be observed, which strongly influences the geometry in two ways: On the one hand, the N=N bond is activated, see that the distance of the azo moiety is enlarged in comparison to the neutral forms by 0.03–0.05 Å, and, on the other hand, the aromaticity of the ring decreases, see the different alternate C–C distances of the phenyl group. Also significant is that the planarity of the HESAB increases in the anionic forms. For instance, in the neutral form 3 the dihedral angle $\tau_{1-3-4-6}$ is 17.4°, while in 3⁻ it decreases to 6.8°.

UV–Visible Absorption Spectroscopy. Experimentally, the UV/vis spectra of these compounds (with 2-hydroxyethyl moieties, vide supra) are recorded in methanol/buffer (1:1) of different pH values. The experimental maximum absorption peaks of compounds 1–6 at specific pH values measured in buffered aqueous methanol are collected in Table 2, and the molar extinction coefficients in pure methanol solutions are given in Table 3.

Before discussing the assignment of the different spectra based on quantum chemical calculations, it is worth reviewing the spectrum of azobenzene, which has been the subject of extensive studies.^{29–32} In the gas phase, the absorption spectrum is characterized by a weak $n\pi^*$ transition at 2.82 eV (440 nm) followed by a strong $\pi\pi^*$ state peaking at 4.12 eV (301 nm).³³ CC2 calculations deliver values of 2.84 and 4.04 eV, respectively, in very good agreement with the experimental values.²⁹ The employment of TD-DFT in azobenzene and derivatives is, however, not straightforward. Charge transfer (CT) states are typically very much underestimated due to the wrong long-range behavior of the applied standard exchange-correlation functionals.³⁴ Several strategies have emerged to consider long-range effects. For instance, range-separated functionals where the total exchange energy is split into short- and long-range contributions have been developed; examples of these functionals are LC-wB97,³⁵ LC-wB97XD,³⁶ or CAM-B3LYP.³⁷ Another possibility is to use hybrid functionals where the exact Hartree–Fock exchange is modified; an illustration is the PBE0 functional, which contains 25% of exact HF exchange. As expected, pure functionals deliver rather poor values in azobenzene. RI-BP86/aug-TZVP predicts the two transitions of azobenzene at 2.19 and 3.35 eV; that is, both bands are underestimated by more than 0.5 eV.²⁹ The hybrid functional B3LYP red-shifts experimental values to a lesser extent, but it still accounts for errors of 0.3–0.4 eV.^{29,30} The systematic study of Jacquemin and co-workers for solvated azobenzene shows that among a large amount of GGA, meta-GGA, conventional hybrids, and the recently developed range-separated hybrid functionals, the PBE0 and CAM-B3LYP functionals give a quantitative agreement on the spectroscopic state of some selected substituted azobenzenes.^{19b} In combination with solvent via the PCM algorithm, mean absolute errors of 0.14 eV for CT in organic dyes, including azobenzene, have been obtained with the PBE0 functional.^{19b} These results are even better than the ones obtained with range-separated hybrids. Recently, Tozer et al. have investigated the relationship between the excitation energy errors and the spatial overlap between the orbitals involved in the excitation, concluding that errors are large when the overlap is low.³⁸ In CT situations where no overlap between the involved orbitals in the CT state is observed, they strongly recommend the use of range-separated functionals, like the CAM-B3LYP functional.³⁹ In azobenzene and the herein studied HESAB dyes, the orbitals involved in the spectroscopic CT excitation should show a high-overlap due to conjugation over the rings (see Figure 2). In such cases, hybrid functionals with augmented amount of exact exchange, like PBE0, can deliver very accurate values because excitations with local character are also present, in agreement with ref 19b.

In the following, we shall investigate first the unsubstituted HESAB system using different protocols for reference. Table 4 collects the excitation energies, oscillator strengths, and the corresponding assignment of the most important peaks of the absorption spectra of the neutral and ionic HESAB model compounds ($1/1^-$) using RI-CC2, BP86, and PBE0 using different geometries, as specified. Figure 3a shows the experimental spectrum of **1** at pH = 3.52 and at pH = 11.08. At these pH values, the neutral (**1**) or the anionic (1^-) form is expected to be predominant, respectively, because the pK_a of **1** in methanol/buffer is 8.35 (in plain buffer, the pK_a is 7.64). Figure 3b–f displays the spectrum of **1** and 1^- at different levels of theory. In Figure 2, the corresponding frontier orbitals are displayed. As we can see, these are very similar for the neutral and ionic forms. The

HOMO–1 of both anionic and neutral compounds $1/1^-$ corresponds to the n_N , the HOMO is a π orbital delocalized mainly in the azo moiety and the phenol moiety, and the LUMO orbital is the antibonding counterpart.

As in azobenzene, the lowest singlet excitation of **1** corresponds to a weak $n\pi^*$ state, followed by a strong $\pi\pi^*$ state. This is predicted by all of the computational approaches but the intensity of the $n\pi^*$ state is underestimated by all theories. In the following and for the sake of clarity, theoretical predictions will be denoted by eV_t (or nm_t) and experimental data by eV_e (nm_e). The most accurate RI-CC2/TZVP method (Figure 3b) obtain excitations at 2.85 eV_t (435 nm_t) and 3.82 eV_t (325 nm_t) for the S_1 ($n\pi^*$) and S_2 ($\pi\pi^*$) states, respectively. The $\pi\pi^*$ state involves an electronic transition from the HOMO to the LUMO, that is, a CT from the azo-phenol moiety to the azo function (see Figure 2). The S_2 $\pi\pi^*$ state, measured at ca. 3.4 eV_e (360 nm_e), is overestimated with RI-CC2 by 0.4 eV probably due to the exclusion of the solvent. The use of a pure functional leads to a similar but opposite error: RI-BP86 predicts the S_2 $\pi\pi^*$ state at 3.07 eV_t , that is, underestimated by 0.4 eV with respect to the experiment (Table 4 and Figure 3c). This shift is reversed with the hybrid functional PBE0 (see Figure 3d), hence predicting a blue-shifted peak with an error of ca. 0.15 eV. The inclusion of solvent effects, only in the calculation of the excitation energies (Figure 3e) or both in the geometry and in the energies (Figure 3f), leads to values around 3.3 eV_t . The $\pi\pi^*$ state is then only blue-shifted with respect to the experiment by 0.1 eV. Because the differences in both latter procedures are not significant, the prescription PCM-PBE0//PBE0, that is, including the solvent only in the calculation of the energies but on vacuum geometries, seems to be a reasonable one to calculate the spectra of the substituted compounds.

The spectrum of 1^- is characterized by a strong band peaking at 2.7 eV_e (460 nm_e). This band corresponds to the $\pi\pi^*$ state, and depending on the computational approach, it is preceded, followed, or embedded into two low-lying very weak $n\pi^*$ transitions. With RI-CC2, the S_1 corresponds to the intense HOMO→LUMO $\pi\pi^*$ transition, calculated at 2.59 eV_t (479 nm_t) and therefore with an error of ca. 0.1 eV. The S_2 and S_3 are excitations from the n_O and n_N , respectively, to the LUMO, at 2.85 eV_t (435 nm_t) and 2.99 eV_t (415 nm_t). With RI-BP86, the S_3 at 2.68 eV_t (464 nm_t) is responsible for the intense $\pi\pi^*$ band, and the two $n\pi^*$ states are below, clearly underestimated with respect to the RI-CC2 values. Even if the RI-BP86 result for the $\pi\pi^*$ peak is in very close agreement with the experiment, one should keep in mind that solvent effects are not included, and hence this agreement is fortuitous. The use of the hybrid PBE0 functional intercalates the $\pi\pi^*$ state at 2.91 eV_t (426 nm_t) between the $n\pi^*$ states. When taking into account solvent effects without and with solvent during the optimizations (Figure 3e,f and Table 4), this transition is red-shifted, appearing at ca. 2.8 eV_t (~440 nm_t), in reasonable agreement with the experimental absorption maximum.

The description of the $n\pi^*$ states deserves additional attention. It has been previously observed that typical global hybrids such as B3LYP or PBE0 show mean absolute errors similar to those of range-separated functionals, like CAM-B3LYP, in small organic dyes.⁴⁰ In solvated substituted azobenzenes, it has been shown that CAM-B3LYP outperforms other functionals, delivering minimum absolute errors as small as 0.02 eV for excitation energy of the $n\pi^*$ state.³² Acceptable residual discrepancies (<0.2 eV) are obtained with the global hybrid PBE0 in the same

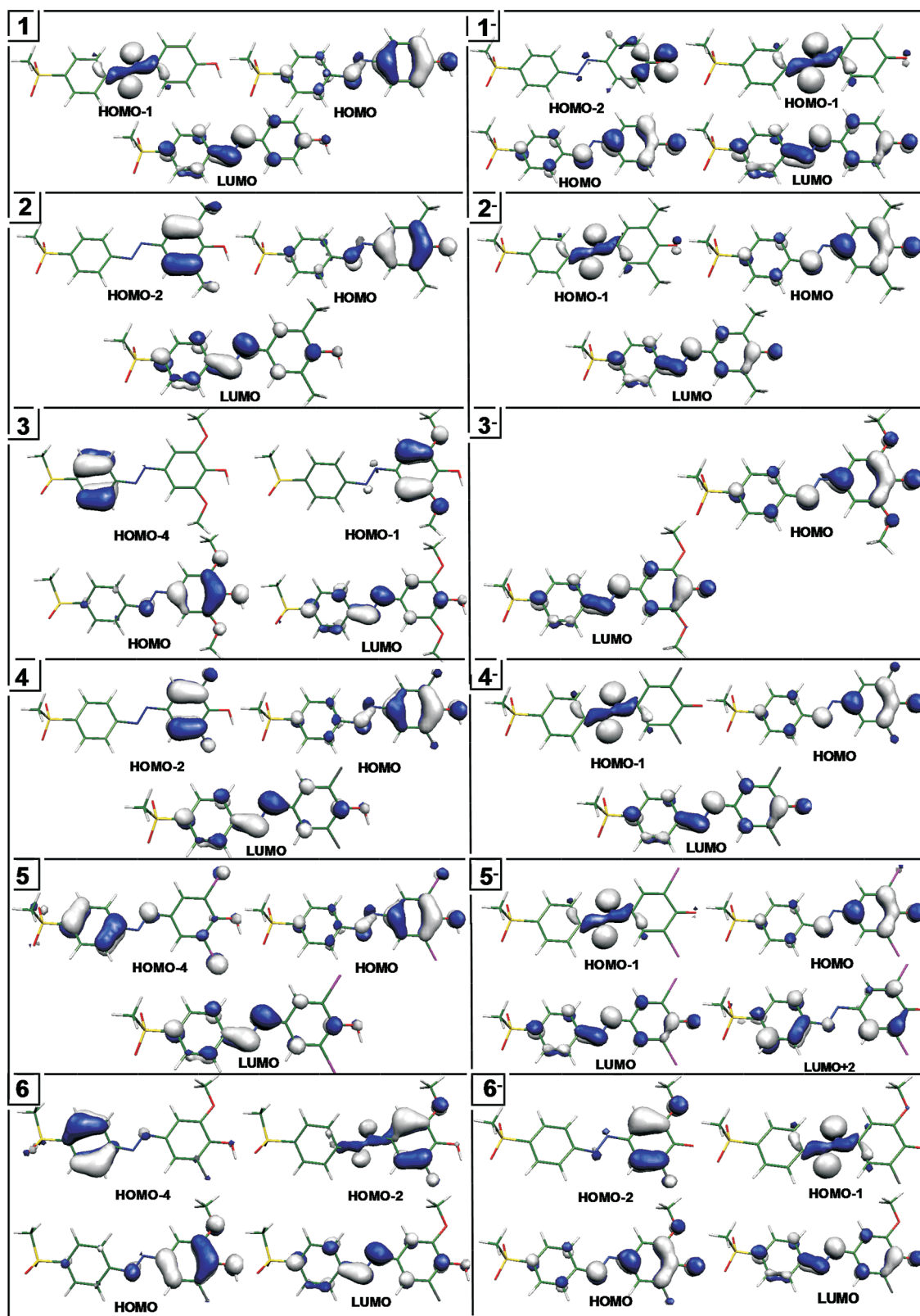


Figure 2. Kohn–Sham frontier orbitals involved in the main electronic transitions of $1/1^-$ to $6/6^-$ HESAB compounds.

compounds.³² In our HESAB compound 1, the $n\pi^*$ state is responsible for the less intense band centered at 2.72 eV_e (456 nm_e); see Table 2. It is likely that the $n\pi^*$ state in compound

1^- is masked under the broad and intense $\pi\pi^*$ band; therefore, we shall limit our discussion to the energies of compound 1. The best agreement for the $n\pi^*$ state is obtained with the accurate RI-

Table 4. Low-Lying Theoretical Electronic Transition Energies, ΔE (in eV and nm), with Oscillator Strengths, f , and Main Assignment (Configuration Interaction Coefficient) for Compounds **1** and Its Corresponding Anion **1⁻**

1					1⁻				
state	ΔE		f	assignment/ c	state	ΔE		f	assignment
	eV	nm				eV	nm		
(RI-CC2//RI-BP86)					(RI-CC2//RI-BP86)				
S ₁	2.85	435	0.017	n _N - π^* (0.92)	S ₁	2.59	479	1.100	π - π^* (0.94)
S ₂	3.82	325	0.999	π - π^* (0.94)	S ₂	2.85	435	0.098	n _O - π^* (0.73)
					S ₃	2.99	415	0.059	n _N - π^* (0.79)
(RI-BP86//RI-BP86)					(RI-BP86//RI-BP86)				
S ₁	2.20	562	0.013	n _N - π^* (0.94)	S ₁	2.01	618	0.001	n _O - π^* (0.95)
S ₂	3.07	403	0.775	π - π^* (0.92)	S ₂	2.15	578	0.010	n _N - π^* (0.94)
					S ₃	2.68	464	1.100	π - π^* (0.92)
(PBE0//PBE0)					(PBE0//PBE0)				
S ₁	2.62	473	0.012	n _N - π^* (0.80)	S ₁	2.65	468	0.086	n _N - π^* (0.57)
S ₂	3.59	345	0.922	π - π^* (0.79)	S ₂	2.91	426	1.210	π - π^* (0.51)
					S ₃	3.05	406	0.008	n _O - π^* (0.68)
(PCM-PBE0//PBE0)					(PCM-PBE0//PBE0)				
S ₁	2.62	473	0.032	n _N - π^* (0.62)	S ₁	2.66	465	0.318	n _N - π^* (0.57)
S ₂	3.29	377	1.027	π - π^* (0.60)	S ₂	2.85	435	1.062	π - π^* (0.51)
					S ₃	3.59	345	0.000	n _O - π^* (0.68)
(PCM-PBE0//PCM-PBE0)					(PCM-PBE0//PCM-PBE0)				
S ₁	2.62	473	0.006	n _N - π^* (0.65)	S ₁	2.71	457	0.002	n _N - π^* (0.66)
S ₂	3.35	370	1.056	π - π^* (0.67)	S ₂	2.82	439	1.318	π - π^* (0.51)
					S ₃	3.58	346	0.000	n _O - π^* (0.68)

CC2 method (2.85 eV_t) and with the global hybrid PBE0 functional (2.62 eV_t). The inclusion of solvent effects on the calculation of the vertical excitations and on the optimized geometries has no effect on the energy of the n π^* excitation (see Table 4). This negligible effect is in agreement with the small shifts that have been observed for other substituted azobenzene dyes when comparing gas-phase and solvated results.³⁰ The intensities of the n π^* bands, however, are different depending on the protocol. The oscillator strengths of the n π^* states of compounds **1** and **1⁻** are higher with PCM-PBE0//PBE0 than with the PBE0/PBE0 model or PCM-PBE0//PCM-PBE0 (see Table 4), and hence PCM-PBE0//PBE0 reproduces best the experimental evidence (see Figure 3e). As a consequence a broader $\pi\pi^*$ band for compound **1⁻** is observed with the PCM-PBE0//PBE0 protocol than with the other approaches (compare Figure 3e with d and f). The increase of n π^* intensity is due to electrostatic effects, because, for instance, the mixing of the n π^* with the $\pi\pi^*$ state for compound **1⁻** is the same with PCM-PBE0//PBE0 and PBE0//PBE0, and thus an increased absorption due to mixing with the strong $\pi\pi^*$ excitation can be ruled out (see wave function coefficients of S₁ in Table 4). A more likely explanation roots to the geometry of the compounds. As we stated above, when considering bulk solvent effects on the optimization of the geometries, slightly more planar optimized structures are obtained. This planarization affects strongly the intensity of the n π^* states of **1** and **1⁻** (compare exemplarily the S₁ oscillator strength values of **1⁻** with the PCM-PBE0//PBE0 and PCM-PBE0//PCM-PBE0 theoretical models). For the latter procedure, a purer n π^* state is obtained, which might contribute to the enormous decreasing of the intensity. Focusing only on the intensities of the n π^* state, it is surprising that the agreement with the experiment is better with PCM-PBE0//PBE0 rather than with the PCM-PBE0//PCM-PBE0 model. A plausible explanation for this fact might be the flat nature of the potential

energy surface in the vicinity of the planarization region, as suggested by the geometric controversy with the different theoretical approaches. To address this geometrical effect and its consequent effects on the intensities of the n π^* state, molecular dynamic simulations evaluating the temperature effects on the UV/vis electronic spectrum might be appropriate, as has been done in *trans*-stilbene.⁴¹

Summarizing, the experimental spectra (Figure 3a) agree reasonably well with those obtained with PCM-PBE0//PBE0 (Figure 3e). The main band of **1** is theoretically blue-shifted by ca. 15 nm, and the one of **1⁻** is red-shifted by 25 nm (see Table 4). Yet, it is fair to mention that the theoretical peaks of **1** and **1⁻** at 377 and 435 nm_t agree much better with the peaks obtained in pure methanol at 364 and 441 nm_e (see Table 3). Therefore, we can infer that the differences between theory and experiment are mainly due to the fact that while the experimental data in Table 4 and Figure 3 are obtained in a buffered 1:1 water/methanol solutions, theory considers only pure methanol as a solvent.

As a general remark, we can conclude that in both the experimental and the theoretical spectra, a strong red (bathochromic) shift can be observed in the band of the anionic form of the dye with respect to the neutral form. This shift can be easily rationalized looking at the responsible transitions in both neutral and anionic forms (cf., Table 4). The involved orbitals, the HOMO/LUMO of the neutral and the anion, exhibit a similar conjugated π character but a different electronic redistribution (Figure 2). Thus, the bands in **1** and **1⁻** have origin in frontier orbitals of similar character but with different energies. Because of additional electrostatic repulsion with the negative charge in **1⁻**, all occupied orbitals are shifted to higher energies, leading to smaller occupied-virtual gaps with respect to **1**, and therefore red-shifted peaks are obtained for the HOMO→LUMO $\pi\pi^*$ transitions. This is responsible for the

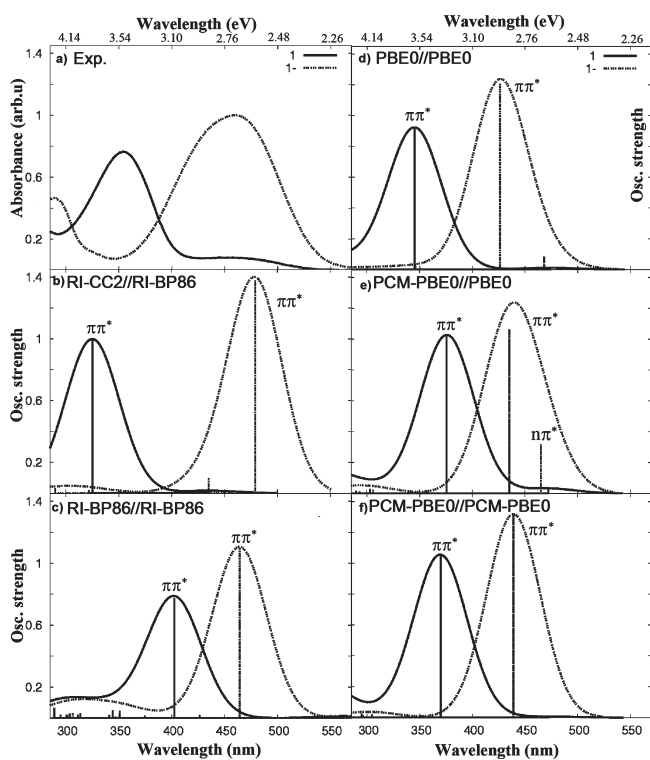


Figure 3. UV-vis absorption spectra of **1** (solid) and **1⁻** (dashed). (a) Experimental spectrum normalized to **1** in arbitrary units, (b) RI-CC2//RI-PB86, (c) RI-BP86//RI-BP86, (d) PBE0//PBE0, (e) PCM-PBE0//PBE0, and (f) PCM-PBE0//PCM-PBE0. The theoretical spectra are convoluted with a Lorentzian function with a full width at half-maximum (fwhm) of 60 nm; the corresponding transitions are marked with vertical lines.

different color of both forms (neutral and anionic) and consequently allows for this compound being used as an indicator (see Table 2).

To conclude, because the spectroscopic states ($\pi\pi^*$), as well as the $n\pi^*$ states of the pair **1/1⁻** are well described with the PCM-PBE0/PBE0 model, this level of theory will be employed to study the HESAB compounds **2/2⁻**–**6/6⁻**.

Substitution Effects in the UV-Vis Spectra. Having analyzed the basic spectral properties of the neutral and anionic forms of the reference compound **1** with respect to the bare azobenzene, we now proceed to evaluate how the electronic transitions are modulated by substitution on the phenol moiety. The corresponding spectra for the HESAB species with electron donor (**2/2⁻** and **3/3⁻**), electron acceptor (**4/4⁻** and **5/5⁻**), and both electron-donor and electron-acceptor (**6/6⁻**) substituents are shown in Figure 4. The main electronic transitions are collected in Table 5, and selected orbitals are found in Figure 2.

As stated above, the substituents influence not only the pK_a value¹⁶ but also the spectral properties, altering the isosbestic point that appears between both anionic and neutral forms in the UV-vis spectrum, and in turn determining to which extent these compounds can be used as indicators. Even if both methyl and methoxy groups donate charge to the aromatic ring, the spectra of **2/2⁻** and **3/3⁻** are different (compare Figure 4a and b). The maximum absorption peaks of **2/2⁻** are separated by 126 nm, while this gap is considerably reduced in the pair **3/3⁻**, showing partially overlapping bands and an isosbestic point less defined.

The main changes are in the spectra of **2** and **3**, while the spectra of **2⁻** and **3⁻** are only shifted by 26 nm. In the dimethyl compound (**2**), the bright S_2 state located at 374 nm corresponds to an HOMO→LUMO $\pi\pi^*$ excitation from the azo and phenol moieties to the π^* orbital, this transition being analogous to the one found in **1**. Additionally, a less intense $\pi\pi^*$ transition contributes to the tail of the band at higher energies (S_3 state, see Table 5). In the dimethoxy compound (**3**), on the other hand, at least two $\pi\pi^*$ states (S_2 and S_3) contribute to the broad band peaking at 460 nm. Presumably the wide profile of the band is due to the S_3 state, which theoretically is determined at 362 nm. An additional intense transition corresponding to the S_5 contributes to the near-UV spectrum; this is an HOMO–4→LUMO excitation from a π orbital localized in the phenyl moiety to π^*_{azo} (see Figure 2). Note that in the case of **3** there is a stabilization of the n_N orbital (HOMO–2). This different behavior obeys the electronic effects: the methoxy group is more electron-donating than the methyl group and hence destabilizes to a major extent the localized and delocalized π orbitals, HOMO–1 and HOMO, respectively, of **3**. Accordingly, these orbitals correspond to HOMO–2 and HOMO, in compound **2**. Consequently, the HOMO→LUMO, $\pi\pi^*_{\text{azo}}$, transition at 403 nm is red-shifted with respect to **2** (compare with 374 nm in **2**).

The agreement between the experimental and theoretical spectra for the pairs **2/2⁻** and **3/3⁻** is reasonable, with general blue-shifts from theory to experiment (Figure 4). The spectrum of **2** is simpler and similar to that of the unsubstituted compound **1**. Likewise, the spectra of the anionic compounds (**2⁻** and **3⁻**) resemble the spectra of **1⁻**. They are both characterized by a strong peak, experimentally located at 499 and 525 nm, and theoretically blue-shifted at 456 nm and 484 nm, respectively. In **2⁻**, both S_1 and S_2 states contribute almost equally to the main band, which then can be assigned as a mixture of the $\pi\pi^*_{\text{azo}}$ and $n\pi^*_{\text{azo}}$ electronic excitations (Table 4). On the contrary, in compound **3⁻**, this peak is due to the S_1 state, which is a HOMO→LUMO transition, localized in the $\pi\pi^*_{\text{azo}}$ orbitals, as in **1⁻**.

The $n\pi^*$ transitions in **2/2⁻** and **3/3⁻** are very similar to those encountered in the unsubstituted pair **1/1⁻** and show similar trends in terms of oscillator strengths (rather small in **2**, **3**, and **3⁻**). As in the pair **1/1⁻**, the $n\pi^*$ transitions are red-shifted upon deprotonation. These effects will also appear in the electron-withdrawing substituted compounds (vide infra). Experimentally, the band of **2** shows a shoulder centered at 427 nm, which is most likely due to the $n\pi^*$ transition (466 nm). The intensity of this transition is theoretically underestimated. Interestingly, in **2⁻**, where the $n\pi^*$ transition is mixed with the $\pi\pi^*$ transition (recall S_1 and S_2 in Table 5), higher oscillator strengths are obtained due to this mixing. To investigate whether this mixing changes when solvent effects are included in the optimization, which redounds in a more planar geometry (vide supra), and whether energies can be improved, test calculations in **2⁻** have been performed with the PCM-PBE0//PCM-PBE0 approach. The resulting energies and oscillator strengths are also included in Table 5 in brackets. As we can see, the mixing between the $n\pi^*$ and $\pi\pi^*$ transitions is slightly reduced (as we saw in the pair **1/1⁻**), rendering different oscillator strengths: a smaller/larger value for the S_1/S_2 states, respectively. However, the solvent effect does not improve substantially the energies, although a slight shift toward the experimental values is obtained.

The neutral forms **4** and **5** show a broad band peaking at 355 and 354 nm, respectively (cf., Table 2). As in the previous cases,

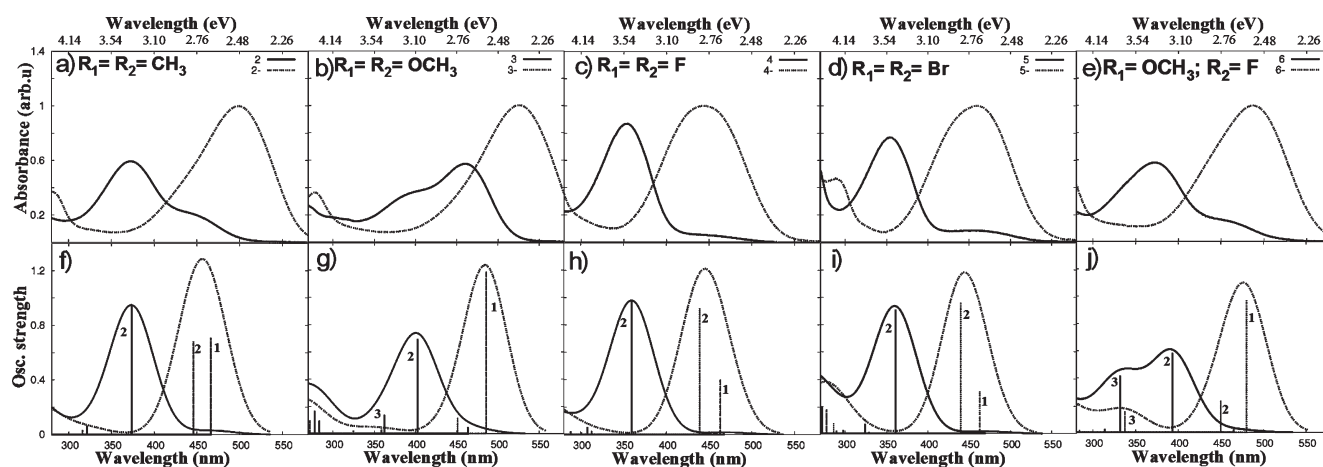


Figure 4. Experimental (top) and PCM-PBE0//PBE0 (bottom) spectra of neutral (solid) and anionic (dashed) compounds. The theoretical spectra are convoluted with a Lorentzian function with a full width at half-maximum (fwhm) of 60 nm; the transitions are marked with vertical lines indicating the corresponding singlet excited state.

Table 5. PCM-PBE0//PBE0 Calculated Electronic Excited States of Compounds 2–6 and Corresponding Anions^a

state	ΔE	f	assignment	state	ΔE	f	assignment
2							
S ₁	2.66(466)	0.029	H-1→L (0.62) n _N -π*	S ₁	2.66(466) [2.71(457)] ^b	0.709 [0.357]	H-1→L (-0.45) n _N -π* H→L (0.45) π-π*
S ₂	3.32(374)	0.940	H→L (0.61) π-π*	S ₂	2.78(446) [2.73(454)] ^b	0.684 [0.988]	[H-1→L (0.57) n _N -π*] [H→L (-0.31) π-π*]
S ₃	3.86(321)	0.063	H-2→L (0.68) π-π*	S ₄	3.86(321) [3.83(324)] ^b	0.023 [0.025]	H-1→L (0.49) n _N -π* H→L (0.41) π-π*
3							
S ₁	2.67(464)	0.048	H-2→L (0.61) n _N -π*	S ₁	2.55(486)	1.190	H-3→L (0.68) π-π* [H-3→L (0.68) π-π*]
S ₂	3.08(403)	0.694	H→L (0.61) π-π*	S ₂	2.75(451)	0.115	H→L (0.57) π-π*
S ₃	3.42(362)	0.138	H-1→L (0.64) π-π*	S ₃	3.46(359)	0.035	H-1→L (0.65) n _N -π*
S ₅	4.47(278)	0.168	H-4→L(0.48) π-π*				
4							
S ₁	2.65(467)	0.016	H-1→L (0.63) n _N -π*	S ₁	2.68(463)	0.404	H-2→L (0.68) π-π*
S ₂	3.45(360)	0.972	H→L (0.62) π-π*	S ₂	2.82(439)	0.921	H-1→L (0.54) n _N -π*
S ₃	4.03(308)	0.052	H-2→L (0.68) π-π*	S ₄	3.97(313)	0.024	H→L (0.49) π-π*
5							
S ₁	2.64(469)	0.016	H-1→L (0.62) n _N -π*	S ₁	2.68(463)	0.314	H-2→L (0.68) π-π*
S ₂	3.44(361)	0.909	H→L (0.62) π-π*	S ₂	2.82(440)	0.959	H-1→L (0.57) n _N -π*
S ₃	3.83(324)	0.072	H-2→L (0.67) π-π*	S ₅	4.17(297)	0.027	H→L (0.53) π-π*
S ₅	4.57(271)	0.198	H-4→L (0.65) π-π*	S ₆	4.34(286)	0.075	H→L+1 (0.63) π-π*
6							
S ₁	2.67(465)	0.022	H-1→L (0.58) n _N -π*	S ₁	2.58(480)	0.973	H→L+2 (0.58) π-π*
S ₂	3.15(393)	0.583	H→L(0.63) π-π*	S ₂	2.75(450)	0.232	H-1→L (0.55) π-π*
S ₃	3.73(332)	0.418	H-2→L (0.61) π-π*	S ₃	3.67(337)	0.154	H-1→L (0.61) n _N -π*
S ₅	4.56(271)	0.116	H-4→L (0.47) π-π*	S ₅	3.95(314)	0.028	H-2→L (0.67) π-π*
6							
H→L+1 (0.66) π-π*							

^a The transitions energies, ΔE , for the most relevant transitions are given in eV (nm) with oscillator strengths, f , and main assignments (configuration interaction coefficient). ^b Values in brackets obtained at the PCM-PBE0//PCM-PBE0 level of theory.

the transitions underlying these bands are the HOMO→LUMO transitions of $\pi\pi^*$ character. In **4**, this band is assigned to the bright S₂ state, located at 360 nm_t, in excellent agreement with the experiment. In **5**, not only the S₂ (361 nm_t) but to a minor extent the much weaker S₃ (324 nm_t) state contributes to this

band. In both compounds, the S₂ state is an HOMO→LUMO transition with $\pi\pi^*$ _{azo} character. In **5**, the S₃ state is an excitation from the HOMO-2, that is, a π orbital localized on the phenol moiety but not on the azo moiety. As it can be seen, the S₂ transition in the electron-withdrawing compounds is blue-shifted

Table 6. Experimental and PCM-PBE0//PBE0 Calculated Ratios between the Heights of the Anion and Neutral Main Bands of Compounds 1–6 and 1[−]–6[−], as well as the Ratio between the Calculated Oscillator Strengths *f* of the Main $\pi\pi^*$ Transitions Underlying the Main Band of the Spectra

	1 [−] /1	2 [−] /2	3 [−] /3	4 [−] /4	5 [−] /5	6 [−] /6
experimental	1.30	1.68	1.74	1.14	1.30	1.72
theoretical	1.21	1.36	1.67	1.24	1.26	1.80
$f^-(\pi\pi^*)/f(\pi\pi^*)$	1.04	1.48 ^a	1.71	0.95	1.06	1.67 ^b

^a Obtained considering the sum of the oscillator strengths of S₁ and S₂ of 2[−], because the $\pi\pi^*$ and $n\pi^*$ are strongly mixed (see Table 5).

^b Obtained considering only the oscillator strength of S₂ of 6.

both theoretically and experimentally in comparison to its analogous S₂ in the electron-donating compounds (2, 3) and the unsubstituted compound 1 (Table 5). This blue-shift can be attributed to the stabilization/destabilization of the HOMO/LUMO pair of orbitals due to electronic effects. Also well represented is the onset of the second band observed in the spectrum of 5 (Figure 4d), which can be assigned to the S₅ state, located at 271 nm_e, and corresponds to an excitation from a π orbital localized on the phenyl ring (HOMO−4, in Figure 2) to the π^*_{azo} orbital.

As with the electron donor derivatives (2[−] and 3[−]) and the bare compound 1[−], the HOMO→LUMO transition is responsible for the electronic properties of 4[−] and 5[−]. Small blue-shifts are observed with respect to 1[−]. As it was found in the other HESAB compounds, the red-shifted bands of 4[−] and 5[−] are more intense than the corresponding neutral ones. We note that the theoretical values are very close to the experimental ones: the theoretical absorption maxima in 5[−] are blue-shifted ca. 12 nm_e and in 4[−] ca. 1 nm_e, with respect to the experiment (cf., Table 3). The main band of 5[−], peaking at 460 nm_e, is explained by the S₁ and S₂ states, which correspond to the $n\pi^*_{\text{azo}}$ and a $\pi\pi^*$ transition, respectively (see Table 5). Additionally, compound 5[−] also shows a weaker band at higher energies (292 nm_e), which is theoretically well described by the S₅ and S₆ $\pi\pi^*$ states (297 and 286 nm_e). It is gratifying to see that in most of the cases, the deviation from the experiment is below 0.1 eV. The maximum discrepancy between theory and experiment is found in the description of the main band of 3, with an error accounting to ca. 0.35 eV. These errors are in the upper limit of accuracy that can be expected for this methodology.

In summary, when comparing the spectroscopic properties of electron-donor and electron-withdrawing substituted HESABs with 1/1[−], we can state the following: First, there is a significant red-shift of the main peaks of the anionic forms of the electron-donor compounds with respect to 1[−] (compare 460 nm_e in 1[−] with 499 and 525 nm_e in 2[−] and 3[−], respectively, in Table 2) due to electronic effects, and although less striking, a small red-shift is present in the main peak of the acid forms (see 361 nm_e in 1 versus 373 and 460 nm_e in 2 and 3). In electron-withdrawing compounds, the trend is inverse; that is, the absorption of the basic forms is slightly blue-shifted in comparison to 1[−] (compare 460 nm_e in 1[−] with 444 and 460 nm_e in 4[−] and 5[−]). The main peaks of the acid forms are also slightly blue-shifted (see Table 2). Despite being small, these differences are recovered by the DFT calculations.

The spectra of the fluoromethoxy pair 6/6[−] (Figure 4e,j) resemble those of 1/1[−] due to compensating electronic effects.

The peak of 6 located at 374 nm_e can be assigned to the S₂ and S₃ states, located at 393 and 332 nm_e. The S₂ state corresponds to the usual HOMO→LUMO transition of $\pi\pi^*_{\text{azo}}$ character. The S₃ is also a $\pi\pi^*$ transition, but starts from the HOMO−2 orbital, which is mainly located in the phenol moiety (see Figure 2), as it has been found in 1, 2, and 4. The spectrum of 6[−] peaks at 487 nm_e and it can be described by the S₁ (480 nm_e), which is the $\pi\pi^*_{\text{azo}}$ transition and a non-negligible contribution of the S₂ $n\pi^*_{\text{azo}}$ state.

Finally, we have considered it of interest to analyze the effect of the pH and substitution pattern on the relative transition intensities. In Table 6, we have calculated the ratio between the heights of the anion and neutral main bands, using the experimental and computed PCM-PBE0//PBE0 spectra. The experimental values show that electron-donor species increases this ratio, while electron-withdrawing groups leave this value almost unchanged. It is gratifying to see that these trends are also theoretically reproduced. Additionally, we have calculated the ratio between the calculated oscillator strengths *f* of the strong $\pi\pi^*$ transitions underlying the peaks of the anion and neutral spectra. The comparison of these ratios indicates the differences in intensity with respect to the reference pair 1/1[−]. As we can see, the trends for electron-donor (2/2[−] and 3/3[−]) and electron-withdrawing (4/4[−] and 5/5[−]) substitution are maintained. The value obtained for the compounds 6/6[−] is not really instructive because the spectrum is broad due to several transitions, but for the sake of uniformity only one transition (S₂) has been taken into account.

5. CONCLUSIONS

In the present Article, the absorption spectra of substituted 2-hydroxyethylsulfonyle azobenzene (HESAB) pH indicator dyes are reported and theoretically assigned with the help of quantum chemical calculations. HESAB indicator dyes can be used for optically monitoring pH in the range from 3 to 10 and can be covalently linked to sensor layers exhibiting high chemical stability, as we have recently reported.¹⁶ HESAB chemistry is not just limited to absorbance spectroscopy but could also be used to develop emission dyes. The absorbance spectra of all the neutral and anionic counterparts are well separated; accordingly, the color changes of HESAB dyes in going from acid to base form (i.e. protonated to deprotonated form) are from yellow to orange or red, hence making feasible its use as pH indicator dyes. Substitution of HESAB complexes by electron-donor and electron-withdrawing moieties biases not only the pK_a values but also the spectroscopic properties of HESAB complexes. The differences of the spectroscopic features upon substitution as well as between the anionic and neutral forms (measured under different pH conditions) have been theoretically elucidated. The good agreement between theory and experiment has been achieved using the density functional protocol PCM-PBE0/PBE0, which includes the solvent effect in the energies with a continuum model but geometries optimized in gas phase. With this theoretical model, deviation from the experiment in the description of the $n\pi^*$ and $\pi\pi^*$ states is below 0.1 eV. Inclusion of additional solvent effects in the optimization of the geometries leads to minor improvements on the transition energies. However, because the geometries optimized in the presence of solvent are more planar than those in gas phase, the intensities of the $n\pi^*$ transitions decrease substantially in comparison to that obtained in gas phase. In the species studied here, this effect leads to a

worse agreement with the weak experimental $n\pi^*$ bands, very likely because temperature effects prevent the molecule from remaining planar, as indicated by the low frequency modes corresponding to the out-of-plane motion of the rings. Whether inclusion of solvent effects in the geometries is necessary in other cases needs to be cautiously evaluated for each particular case.

In general, from this study, the following conclusions are extracted for the HESAB dyes:

- (i) The spectroscopic state of all of the neutral and anionic HESAB dyes here investigated is the $\pi\pi^*$ transition. This is then the state that determines the functionality of these complexes as pH indicators, while the $n\pi^*$ transitions are much weaker or even dark.
- (ii) In all of the HESAB dyes, a red-shift is observed upon deprotonation. This effect can be trivially explained in terms of the additional electrostatic repulsion between the negative charge and the occupied orbitals, which are then shifted to higher energies, thus leading to smaller occupied-virtual gaps.
- (iii) In the electron-donor compounds, both the neutral and the anionic forms show peaks red-shifted with respect to the unsubstituted compound. An inverse trend is observed in the studied electron-withdrawing derivatives; peaks are blue-shifted with respect to the unsubstituted compound. These effects have been rationalized theoretically in terms of the stabilization/destabilization of the orbital levels due to electronic effects upon substitution. The consequences of substitution on the optical properties of the pH indicators are translated in a change of the color pattern between the pair of complexes, for instance, going from pale yellow to orange in $1/1^-$ and from yellow/orange to red in $3/3^-$.
- (iv) In the compound where both electron-donor and -withdrawing substituents are present, the spectrum resembles very much that of the unsubstituted HESAB, due to compensating electronic effects.

When planning the synthesis of new indicator dyes, one important issue is the possible prediction of their future optical properties. This is important when sensor dyes have to be made spectrally compatible with cheap light sources (e.g., light emitting diodes or laser diodes). Furthermore, well-separated absorbance spectra between acid and base form simplify the setup of the optical sensor device and enhance the signal-to-noise ratio. In the present work, a good correlation between calculated and experimental absorbance spectra has been achieved, thus paving the way for the dedicated design of new sensor dyes and sensor devices.

AUTHOR INFORMATION

Corresponding Author

*Fax: +49 3641 948302. E-mail: leticia.gonzalez@uni-jena.de.

ACKNOWLEDGMENT

This work has been funded by the Carl-Zeiss foundation (D.E.), the BMBF project "Aquaoptrode" (no. 13N9535), the projects MO 1062/5-1 and MO 1062/6-1 of the Deutsche Forschungsgemeinschaft, and the project AZ-Nr.: 20.10-3410-2 (Projekt Sensormaterialien) of the Bayerische Staatsministerium für Wirtschaft, Infrastruktur, Verkehr und Technologie.

Computer time in the Rechenzentrum of the Friedrich-Schiller-Universität Jena is gratefully acknowledged.

REFERENCES

- (1) McMillan, G. K.; Cameron, R. A. *Advanced pH Measurement and Control*, 3rd ed.; ISA: Research Triangle Park, NC, 2004.
- (2) See, for example: (a) Orellana, G., Moreno-Bondi, M. C., Eds. *Springer Series on Chemical Sensors and Biosensors*; Springer-Verlag: Berlin, Germany, 2004; Vol. 1. (b) Narayanaswamy, R., Wolfbeis, O. S., Eds. *Optical Sensors: Industrial, Environmental and Diagnostic Applications*; Springer: Berlin, Heidelberg, 2004. (c) Wolfbeis, O. S. *Anal. Chem.* **2006**, *78*, 3859.
- (3) Trupp, S.; Alberti, M.; Carofiglio, T.; Lubian, E.; Lehmann, H.; Heuermann, R.; Yacoub-George, E.; Bock, K.; Mohr, G. J. *Sens. Actuators, B* **2010**, *150*, 206.
- (4) (a) Durr, H.; Bouas-Laurent, H. *Photochromism, Molecules and Systems*; Elsevier: Amsterdam, 1990. (b) Barrett, C. J.; Mamiya, J.; Yager, K. G.; Ikeda, T. *Soft Matter* **2007**, *3*, 1249.
- (5) See, for example: Feringa, B. L. *Molecular Switches*; Wiley-VCH: Germany, 2003.
- (6) Shinkai, S.; Manabe, O. *Top. Curr. Chem.* **1984**, *121*, 76.
- (7) Hugel, T.; Holland, N. B.; Cattani, A.; Moroder, L.; Seitz, M.; Gaub, H. E. *Science* **2002**, *296*, 1103.
- (8) Natansohn, A.; Rochon, P. *Chem. Rev.* **2002**, *102*, 4139.
- (9) Spörlein, S.; Carstens, H.; Satzger, H.; Renner, C.; Behrendt, R.; Moroder, L.; Tavan, P.; Zinth, W.; Wachtveitl, J. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 7998.
- (10) Carofiglio, T.; Fregonese, C.; Mohr, G. J.; Rastrelli, F.; Tonellato, U. *Tetrahedron* **2006**, *62*, 1502.
- (11) Mohr, G. J. *Anal. Chim. Acta* **2004**, *508*, 233.
- (12) Mohr, G. J. *Dyes Pigm.* **2004**, *62*, 77.
- (13) Mohr, G. J. *Sens. Actuators, B* **2003**, *90*, 31.
- (14) Ward, C. J.; Patel, P.; James, T. D. *J. Chem. Soc., Perkin Trans. 2002*, *1*, 462.
- (15) Mohr, G. J. *Chem. Commun.* **2002**, *22*, 2646.
- (16) Mohr, G. J.; Müller, H.; Bussemer, B.; Stark, A.; Carofiglio, T.; Trupp, S.; Heuermann, R.; Henkel, T.; Escudero, D.; González, L. *Anal. Bioanal. Chem.* **2008**, *392*, 1411.
- (17) (a) Runge, E.; Gross, E. K. U. *Phys. Rev. Lett.* **1984**, *52*, 997. (b) Stratmann, R. E.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **1998**, *109*, 8218. (c) Casida, M. E. In *Time-Dependent Density-Functional Response Theory for Molecules*; Chong, D. P., Ed.; World Scientific: Singapore, 1995; Vol. 1, pp 155–192. (d) Koch, W.; Holthausen, M. C. *A Chemist's Guide to Density Functional Theory*; Wiley-VCH: Germany, 2000. (e) Barone, V.; Polimeno, A. *Chem. Soc. Rev.* **2007**, *36*, 1724. (f) Jacquemin, D.; Perpète, E. A.; Ciofini, I.; Adamo, C. *Acc. Chem. Res.* **2009**, *42*, 326.
- (18) See, for example: (a) Goerigk, L.; Moellmann, J.; Grimme, S. *Phys. Chem. Chem. Phys.* **2009**, *11*, 4611. (b) Goerigk, L.; Grimme, S. *J. Chem. Phys.* **2010**, *132*, 1841103. (c) Jacquemin, D.; Perpète, E. A.; Ciofini, I.; Adamo, C. *J. Chem. Theory Comput.* **2010**, *6*, 1532.
- (19) (a) Jacquemin, D.; Preat, P.; Wathélet, V.; Fontaine, M.; Perpète, E. A. *J. Am. Chem. Soc.* **2006**, *128*, 2072. (b) Jacquemin, D.; Perpète, E. A.; Scuseria, G. E.; Ciofini, I.; Adamo, C. *J. Chem. Theory Comput.* **2008**, *4*, 123.
- (20) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158.
- (21) (a) Cossi, M.; Barone, V.; Mennucci, B.; Tomasi, J. *J. Chem. Phys. Lett.* **1998**, *286*, 253. (b) Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *106*, 5151.
- (22) Christiansen, O.; Koch, H.; Jørgensen, P. *Chem. Phys. Lett.* **1995**, *243*, 409.
- (23) Hättig, C.; Weigend, F. *J. Chem. Phys.* **1995**, *243*, 409.
- (24) Becke, A. D. *Phys. Rev. A* **1998**, *38*, 3098.
- (25) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822.
- (26) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165.

- (27) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision C.02; Gaussian, Inc.: Wallingford, CT, 2004.
- (28) Bouwstra, J. A.; Schouten, A.; Kroon, J. *Acta Crystallogr., Sect. C* **1983**, *25*, 3561.
- (29) Fliegl, H.; Köhn, A.; Hättig, C.; Ahlrichs, R. *J. Am. Chem. Soc.* **2003**, *125*, 9821.
- (30) Briquet, A.; Vercauteren, D. P.; André, J.-M.; Perpète, E. A.; Jacquemin, D. *Chem. Phys. Lett.* **2007**, *435*, 257.
- (31) Briquet, A.; Vercauteren, D. P.; Perpète, E. A.; Jacquemin, D. *Chem. Phys. Lett.* **2006**, *417*, 190.
- (32) Jacquemin, D.; Perpète, A.; Scuseria, G. E.; Ciofini, I.; Adamo, A. *Chem. Phys. Lett.* **2008**, *465*, 226.
- (33) Andersson, J.-A.; Petterson, R.; Tegner, L. *J. Photochem.* **1982**, *20*, 17.
- (34) Dreuw, A.; Head-Gordon, M. *Chem. Rev.* **2005**, *105*, 4009.
- (35) Chai, J.-D.; Head-Gordon, M. *J. Chem. Phys.* **2008**, *128*, 084106.
- (36) Chai, J.-D.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615.
- (37) Yanai, T.; Tew, D.; Handy, H. *Chem. Phys. Lett.* **2004**, *393*, 51.
- (38) Peach, M. J. G.; Benfield, P.; Helgaker, T.; Tozer, D. J. *Chem. Phys.* **2008**, *128*, 044118.
- (39) Peach, M. J. G.; Ruth Le Sueur, C.; Ruud, K.; Guillaume, M.; Tozer, D. J. *Phys. Chem. Chem. Phys.* **2009**, *11*, 4465.
- (40) Jacquemin, D.; Perpète, E. A.; Vydrov, O. A.; Scuseria, G. E.; Adamo, A. J. *Chem. Phys.* **2007**, *127*, 094102.
- (41) Kwasniewski, S. P.; François, J. P.; Deleuze, M. S. *Int. J. Quantum Chem.* **2001**, *85*, 557.

A Computational Study (TDDFT and RICC2) of the Electronic Spectra of Pyranoanthocyanins in the Gas Phase and Solution

Angelo Domenico Quartarolo and Nino Russo*

Dipartimento di Chimica and Centro di Calcolo ad Alte Prestazioni per Elaborazioni Parallele e Distribuite-Centro di Eccellenza MIUR, Università della Calabria, I-87030 Arcavacata di Rende, Italy

S Supporting Information

ABSTRACT: The conformational structures and UV–vis absorption electronic spectra of a class of derived anthocyanin molecules (pyranoanthocyanins) have been investigated mainly by means of density functional (DFT) and time-dependent DFT methods. Pyranoanthocyanins are natural pigments present in aged wines and absorb at shorter wavelengths (around 500 nm) than the parent anthocyanin compounds, giving an orange-brown colored solution. The investigated molecules are derived from the reaction of glycosylated malvidin, peonidin, and petunidin with enolizable molecules (acetaldehyde and pyruvic acid) and vinyl derivatives. During wine storage, the concentration of pyranoanthocyanins increases with time, and analytical measurements (e.g., UV–vis spectroscopy) can characterize aged wines by color analysis. The prediction of absorption electronic spectra from TDDFT results, with the inclusion of water bulk solvation effects through the conductor-like polarizable continuum model, gives an absolute mean deviation from experimental absorption maxima of 0.1 eV and a good reproduction of the spectra line shape over the visible range of the spectrum. TDDFT calculated excitation energies agree with those obtained from *ab initio* multireference coupled cluster with the resolution of identity approximation (RICC2) methods, calculated at DFT gas-phase geometries.

1. INTRODUCTION

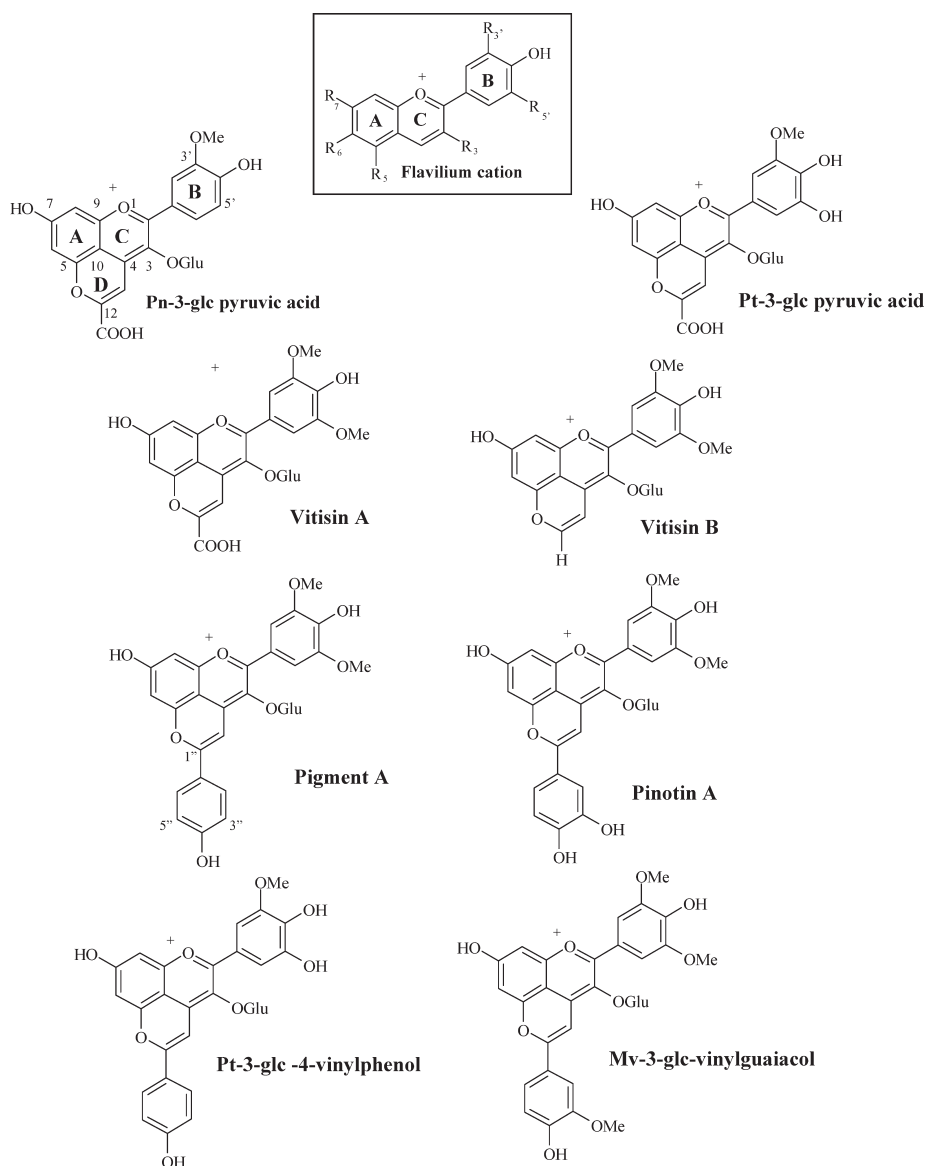
Anthocyanins are naturally occurring pigments present in the tissues of vascular plants, including leaves, flowers, fruits, and roots, and belong to the general class of flavonoids compounds.^{1–4} They are based on the flavilium ion or 2-phenylchromenylium with the basic chemical structure, shown in Scheme 1, constituted by an aromatic ring (A) fused to an oxygen-containing heteroaromatic ring (C) and a third ring (B), which is connected to ring C through a carbon–carbon single bond.⁵ The benzopyrilium ring (A and C fused rings) assumes a rigid planar conformation, while the substituted phenyl group (ring B) can freely rotate, giving rise to different conformational energy minima. Side R groups in Scheme 1 can be represented by hydrogen, hydroxyl, methoxy, glucoside, and its derivatives, and their possible combination gives a great number of different structures.² Depending upon the presence or not of the glycosyl moiety, anthocyanins are respectively named anthocyanin glycosides or anthocyanidin aglycons. They play an important biological role in plant metabolism, for example, in the pollination, reproduction, or photoprotection of the plants against high-energy UV solar radiation damage of plant cells.⁶ This latter function is the basis of the antioxidant activity of anthocyanins to scavenge free radicals produced by metabolism and can also give benefits to human health for the prevention of different diseases like cardiovascular illnesses, diabetes, or tumors.^{7–9} An important feature of anthocyanins is that they are responsible for fruit and flower colors, which differ according to the nature of the anthocyanin pigments present, since visible absorption maxima and intensity are strictly related to their molecular structures.^{10–13} The color property of anthocyanins is also employed in the food industry, where they are used as vegetable colorants together with carotenoid pigments replacing synthetic dyes.¹⁴

The molecular stability to degradation of pure anthocyanin solutions depends on different factors such as the concentration, pH variation, molecular structure, storage temperature, light, complexation with metallic ions in solution, and oxygen.^{11,15,16} The analysis of these factors can be helpful in the food industry for improving the stabilization of anthocyanins as food colorants. The anthocyanin solution concentration, the pH variation, and the presence of copigments and/or metallic ions also affect color properties. In solution, anthocyanins exist in different equilibrium forms depending on pH values. In strong acidic aqueous solutions (pH = 1–3), the predominant species present in solution is the flavilium cation (Scheme 1), which gives a strong red color. Increasing the pH between 2 and 4, the quinoidal species, formed after proton abstraction from the hydroxyl groups, are dominant with a bathochromic wavelength shift. At a pH between 5 and 6, after water hydrolysis, the colorless carbinol pseudobase and chalcone species appear in solution, while at alkaline pH, anthocyanins tend to degrade depending on the nature of the substituent R groups on the B ring. The larger the number of methoxy or hydroxyl groups on the B ring, the less stable the corresponding anthocyanidin is in neutral media. On the other hand, anthocyanin glycosides are more stable since the presence of the glycosyl moiety prevents the degradation reaction.^{1,17} A class of anthocyanin-derived pigments or pyranoanthocyanins (Scheme 1), with different properties from anthocyanins, was detected in red wine filtrates by Cameira dos Santos et al. in 1996 and has attracted much attention in recent years for its possible use in characterizing aged wines.¹⁸ Pyranoanthocyanins are more stable at different pH values with a

Received: December 23, 2010

Published: March 23, 2011

Scheme 1. Molecular Structures with Substituent (R) and Ring Labelling for Flavilium Cation (Upper Panel) and Investigated Pyranoanthocyanins



hypsochromic shift of the absorption wavelength maxima λ_{\max} in comparison with the anthocyanin monoglucosides, and that gives an orange-brown color to their solution.¹⁹ These compounds are derived from the reaction of anthocyanins with low molecular weight molecules as flavonols, pyruvic acid, and 4-vinylphenol; the cyclization that takes place between the carbon at position 4 and the hydroxyl group at position 5 (Scheme 1) forms a fourth ring (ring D in Pe-3-glc pyruvic acid of Scheme 1) or pyran ring.^{20,21} The stability of pyranoanthocyanins is due to the formation of the pyran ring that works as a protective group against the nucleophilic addition of water, avoiding the formation of the colorless carbinol pseudobase and decolorization by bisulphite (SO₂). The formation of pyranoanthocyanins in model solutions has been found to be fast and dependent on the initial concentrations of anthocyanins and the reaction partners (e.g., acetaldehyde, pyruvic acid, and other enolizable molecules) as well as on pH and temperature

conditions.^{22,23} Another important factor for their formation is the storage time; in fact, in red wines, although absent in the initial products, the concentration of pyranoanthocyanins increases with time.²⁴ The occurrence of pyranoanthocyanins was investigated and detected also in fermented and unfermented juices of black carrots, blood oranges, and strawberry fruits.^{25–27} The identification methods for anthocyanins and pyranoanthocyanins are based on high-performance liquid chromatography combined with mass spectrometry and other spectroscopic techniques like nuclear magnetic resonance and UV–vis spectroscopy.^{28–33} The latter is an important analytical tool giving both quantitative and qualitative information, like acid constant determinations, with low cost.^{34,35} In recent years, theoretical studies, both *ab initio* and semiempirical, have also been done, in order to investigate different aspects of the chemistry of anthocyanins and pyranoanthocyanins.^{36–38} Density functional methods (DFT), for example, have been used to

predict the stability of charge transfer complexes between anthocyanins in the presence of copigment molecules (like hydroxycinnamic acids, e.g., gallic or caffeic acid)³⁹ or, in the case of pyranoanthocyanins, to study the reaction mechanisms responsible for the antioxidant activity.^{40,41} Theoretical studies on UV–vis spectra of anthocyanins and pyranoanthocyanins have also been done by using semiempirical methods (e.g., ZINDO) or the time-dependent DFT approach (TDDFT).^{42,44} In this paper, the electronic absorption spectra of pyranoanthocyanins, reported in Scheme 1, will be mainly investigated by the TDDFT methodology, which has become, in the past decade, a well assessed theoretical tool for the simulation of the electronic spectra (absorption, fluorescence, and phosphorescence) of medium and large organic and transition metal containing molecules.⁴⁵ The coupled cluster with approximate singles and doubles method and the resolution of identity approximation (RICC2), a correlated and size-consistent method, will also be applied to pyranoanthocyanins as an *ab initio* multiconfigurational reference approach.⁴⁶ The main purpose of this work is to theoretically characterize the electronic spectra and find a possible structure–property correlation for the pyranoanthocyanins in Scheme 1, together with a comparison of the theoretical results with available experimental electronic spectra.

2. COMPUTATIONAL METHODS

All calculations were carried out with the Turbomole software package on a Quad-Core AMD Opteron processor (2.7 CPU GHz with 62.9 GB memory).⁴⁷ Structures were preoptimized at the DFT theory level with the Becke–Perdew exchange–correlation functional (PBE)^{48,49} and the resolution of identity approximation (RIDFT module)^{50,51} avoiding the direct four-center integral calculation. The double- ζ -quality SV(P) and corresponding auxiliary basis sets, with polarization functions for C and O atoms, were adopted during this step.^{52,53} The final structure optimizations were done with the PBE0 free-parameter hybrid functional,⁵⁴ which adds a fixed fraction (1/4) of the Hartree–Fock exact exchange energy to the PBE exchange–correlation functional,⁵⁵ and the SV(P) basis set. Optimized stationary points were characterized as energy minima by vibrational frequency calculations. An extensive basis set benchmark has been performed on Peonidin-3-glucoside pyruvic acid, in order to assess the basis set influence on the main excitation energy (experimental maximum absorption at 2.47 eV). The Ahlrichs double- [SV(P), SVP with polarization functions for hydrogens]⁵² and triple- ζ basis sets (TZVP)⁵⁶ give an excitation energy of 2.30 eV (539 nm), for both SV(P) and SVP basis sets, and 2.29 eV (540.5 nm) for the TZVP basis set. In the case of the SV(P) basis set, the addition of s and p diffuse functions for carbon and oxygen atoms yields a maximum absorption at 2.29 eV (541.5 nm). The use of correlation-consistent basis sets (cc-pVDZ and cc-pVTZ) developed by Dunning⁵⁷ gives values of 2.31 eV (535.8 nm) and 2.32 eV (534.5 nm), respectively. The maximum absorption difference, between the SV(P) and the more extended triple- ζ basis sets (TZVP and cc-pVTZ), is small and within 0.02 eV (10 nm). On the basis of these results, excitation energies were calculated on PBE0/SV(P) optimized structures by means of the TDDFT method at the same level of theory. This approach has been successfully applied for the prediction of UV–vis electronic spectra of organic and metal-transition containing systems, yielding an error within 0.3–0.4 eV.^{58–61}

Bulk solvent effects were taken into account both on geometries and excitation energies by means of the conductor-like polarizable continuum model (CPCM)^{62,63} by setting the dielectric constant ϵ to 78.39 in order to simulate an aqueous medium, since pyranoanthocyanins are water-soluble. For the cavity construction on each atom, default parameters (solvent radius and optimized atomic radii) as defined in the COSMO module were chosen. The band shapes of the electronic spectra were reproduced, using the SWizard program,⁶⁴ by a sum of Gaussian functions centered on each excitation energy according to the formula

$$\varepsilon(\omega) = 2.174 \times 10^8 \sum_I \frac{f_I}{\Delta_{1/2}} \exp\left(-2.733 \frac{(\omega - \omega_I)^2}{\Delta_{1/2}}\right) \quad (1)$$

where molar absorbance ε is given in $M^{-1} \text{ cm}^{-1}$ units and ω_I and f_I are, respectively, the excitation energies and oscillator strengths for each allowed electronic transition. The sum in eq 1 is such that the total integrated intensity under the absorption profile is equal to the sum of the oscillator strengths f_I . A constant half-height bandwidth $\Delta_{1/2}$ of 0.3 eV has been chosen in order to match the corresponding experimental spectra. The lowest 30 excitation energies were included in the spectra simulation. For a theoretical comparison with TDDFT excitation energies, also the RICC2 model has been applied at gas-phase optimized DFT geometries, with the frozen orbital space option.⁶⁵ Moreover, in order to make computationally feasible the RICC2 calculations, the SV(P) basis set has been employed through all of the calculations. In fact, as found for the TDDFT basis set benchmark, the use of larger basis sets results in little improvement in the accuracy of the main excitation energy. For the Pn-3-glc pyruvic acid derivative, the use of TZVP and correlation-consistent basis sets (cc-pVDZ, cc-pVTZ, and aug-cc-pVDZ) yields the main excitation energy at 2.33 eV [SV(P), SVP, cc-pVDZ] and 2.37 eV (TZVP, cc-pVTZ, and aug-cc-pVDZ). Optimized Cartesian coordinates for all molecular structures (in water) are included in the Supporting Information (pp S2–S13).

3. RESULTS AND DISCUSSION

3.1. Conformational Structures. The investigated pyranoanthocyanins reported in Scheme 1 are structurally derived from the condensation reaction of organic molecules with three kinds of glycosylated anthocyanins: malvidin- ($R_{3'} = R_{5'} = \text{OCH}_3$), peonidin- ($R_{3'} = \text{OCH}_3$, $R_{5'} = \text{H}$), and petunidin-3-glucosides ($R_{3'} = \text{OCH}_3$, $R_{5'} = \text{OH}$), abbreviated in the text respectively as Mv-, Pe-, and Pt-3-glc (see Scheme 1). When the reaction occurs between pyruvic acid and Pn-, Pt-, and Mv-3-glc, the products are the two pyruvic derivatives and Vitisin A, whereas the reaction of Mv-3-glc with acetaldehyde gives the pyranoanthocyanin derivative named Vitisin B. The main structural difference among them is represented by the side ring C substituent groups at positions 3' and 5' and in the case of Vitisin B by the presence of the hydrogen atom at position 12' in place of the carboxylic group. A different structure modification can be obtained from the reaction of anthocyanins with vinylphenol, vinylcatechol, and vinylguaiacol, which causes an elongation of the molecular system at position 12'. The reaction between Mv- and Pt-3-glc with 4-vinylphenol gives, respectively, Pigment A and Pt-3-glucoside-4-vinylphenol derivatives, while Pinotin A and Mv-3-glc-vinylguaiacol come from the reaction between Mv-3-glc and vinylcatechol and vinylguaiacol molecules, respectively

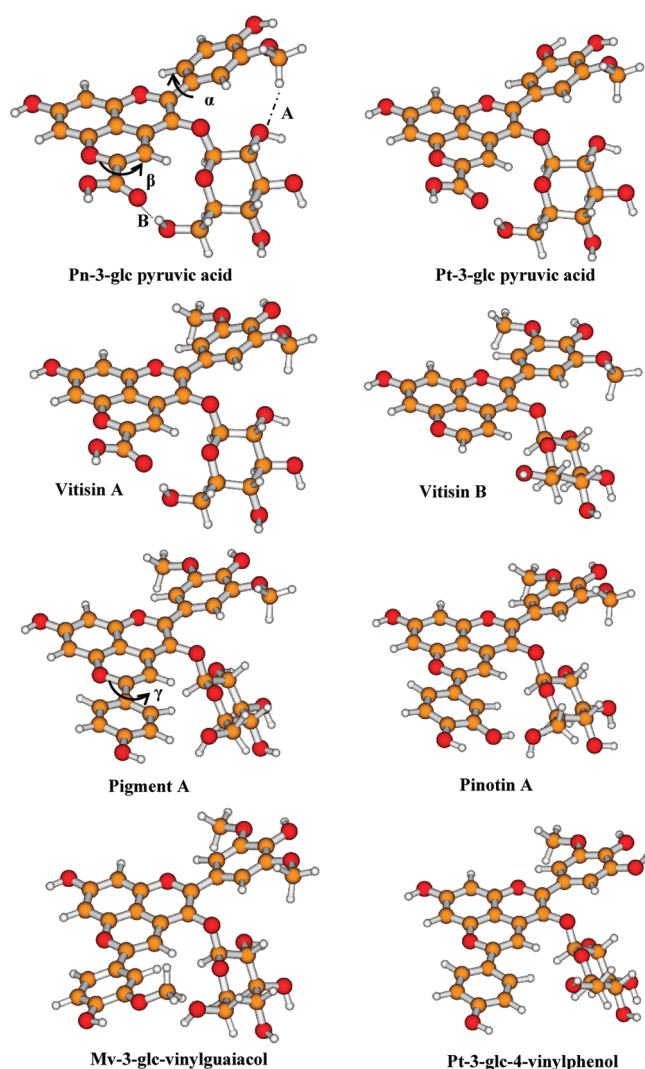


Figure 1. Pyranoanthocyanin global minimum structures.

(Scheme 1). For all studied pyranoanthocyanins, only the red flavilium cation was taken into account for molecular properties calculations. In fact, pyranoanthocyanins are less sensitive to pH increases with respect to anthocyanins, so for example at wine pH (about 3.6), the dominant contributing molecular structure, in solution equilibrium, can be considered the flavilium cation. Like anthocyanins, pyranoanthocyanins can coexist in solution in different conformational structures because of the free rotation around carbon–carbon single bonds. In particular, three main conformational degrees can be identified: (a) the rotation of ring B around the carbon–carbon single bond that connects it to ring C (torsional angle α in Figure 1), (b) the rotation around the bond connecting the carbon atom at position 12 of Scheme 1 to carboxylic (angle β in Pn-3-, Pt-3-glc derivatives and Vitisin B in Figure 1) and substituted phenyl groups (angle γ in Pigment A and the analogous parameter in vinyl derivatives in Figure 1), and (c) the free rotation of hydroxyl and methoxy groups attached at different ring positions (A–D rings in Scheme 1) and a sugar moiety, which can form intramolecular stabilizing hydrogen bonds. The sugar moiety has been chosen as in the β -D-glucopyranose conformation, having the hydroxymethyl group oriented in the equatorial (gauche) position, which is the most

abundant anomeric form in water.^{66,67} The optimized structures of the most stable conformers for each pyranoanthocyanin are shown in Figure 1, while that of the less stable conformers and the relative energy (kcal/mol) with respect to the most stable conformer are reported in the Supporting Information (pp S14–S16).

The energetic differences are within 10 kcal/mol and are in some cases negligible. The driving force that stabilizes the structures in Figure 1, or in general each conformer to the less stable one, can be attributed to the number of hydrogen bonds formed. The value of the dihedral angle α is due to a delicate balance between the degree of electronic delocalization (through ring B and the rest of the molecule) and the nature of the interactions of ring B hydrogens with the glycosyl group. For malvidin-based pyranoanthocyanins (e.g., Vitisin-A and Pigment A), the methyl groups on ring C are oriented in opposite directions, so the main conformational variable is represented by the group orientation at position 12' (carboxylic or vinyl derivatives). Pyruvic acid adducts and Vitisin-B have the C ring rotated by about 20° (dihedral angle α in Figure 1) and the carboxylic group rotated by about 15–17° (dihedral angle β in part A). The oxygen atom forms a hydrogen bond with a length between 1.8 and 1.9 Å (interaction B in Figure 1). For Vitisin B, the absence of the carboxylic group reduces the steric interaction between the sugar moiety and ring C, and as a consequence, there is a decrease in angle α (about 9°). A weaker interaction (about 2.4 Å) exists between the methyl group oxygen on the C ring and the sugar hydroxyl group (A interaction in Figure 1). For vinyl derivatives (Pigment A, Pinotin A, Mv-3-glc, and Pt-3-glc derivatives), angle α is similar to that found for pyruvic and acetaldehyde adducts, and it ranges between 16 and 20°. The torsional angle γ (see Pigment A and other vinyl derivatives in Figure 1) that describes the mutual rotation between ring D and substituted phenyl is small for Pigment A and Pt-3-glc-4-vinylphenol (respectively about 4 and 6°). For these two molecules, the phenyl group has one hydroxyl substituent at the *para* position that does not interact, through a hydrogen bond, with the glycoside moiety. The most stable conformers of Pinotin A and Mv-3-glc-4-vinylguaiaicol are characterized by a dihedral angle γ of 12 and 19°, respectively. The increased γ rotation is mainly due to the formation of hydrogen bonding between the methoxy (for Pinotin A) and hydroxyl (for Mv-3-glc-4-vinylguaiaicol) substituents at position 3'' and the $-\text{CH}_2\text{OH}$ (hydroxymethyl) group on the glycoside moiety.

3.2. Electronic Spectra. The TDDFT results, regarding *in vacuo* and water (CPCM model) excitation energies (in eV and nm units), transition configurations, and oscillator strengths, have been collected in Tables 1 and 2. In particular, in Table 1, pyruvic and acetaldehyde anthocyanin derivatives are grouped together, whereas in Table 2, the results obtained for the vinyl derivatives that possess a structure elongation at position 12 of ring D (see Scheme 1) are shown. For all three pyruvic and acetaldehyde adducts in Table 1, the most intense signal is given by a HOMO to LUMO transition ($\pi \rightarrow \pi^*$ with charge transfer character, see Figure 2), with oscillator strengths f between about 0.3 and 0.4. For Vitisin A, in the visible region, there are two *in vacuo* excitation energies with comparable strengths: the first at 573 nm ($f = 0.2347$) and the second at 506 nm ($f = 0.2834$). In this case, the convoluted simulated spectrum gives the maximum absorption band centered at 529 nm. In the same way, for Vitisin B, the maximum absorption wavelength (λ_{max}) is at 488 nm, which is lower by 10 nm than the first excitation energy (2.49 eV,

Table 1. Calculated Excitation Energies ΔE (eV, nm), Main Configuration (Percentage Contribution in Parentheses), and Oscillator Strengths f (*in vacuo* and Water), for Peonidin- and Petunidin-3-Glucoside Pyruvic Acids (Pn-3-glc and Pt-3-glc der.), Vitisin A, and Vitisin B^a

molecule	state	TDDFT ^b						exptl.
		vacuum			c-pcm (water)			
		ΔE (eV,nm)	f	configuration	ΔE (eV,nm)	f		
Pn-3-glc der.	1 ¹ A	2.30, 539	0.2971	H \rightarrow L (94.5)	2.45, 506	0.3063	2.47, 503 ^c	
	2 ¹ A	2.81, 442	0.1970	H-1 \rightarrow L (93.8)	2.85, 435	0.1235		
	3 ¹ A	2.89, 429	0.0350	H-2 \rightarrow L (97.3)	3.16, 392	0.0654		
Pt-3-glc der.	1 ¹ A	2.27, 547	0.0084	H-1 \rightarrow L (96.6)	2.47, 503	0.4195	2.45, 507 ^c	
	2 ¹ A	2.35, 527	0.4737	H \rightarrow L (95.5)	2.52, 492	0.0486		
	3 ¹ A	2.87, 432	0.0116	H-2 \rightarrow L (86.4)	2.95, 421	0.0138		
Vitisin A	1 ¹ A	2.16, 573	0.2347	H \rightarrow L (78.3)	2.40, 518	0.3892	2.45, 507 ^c ; 2.44, 509 ^d	
				H-1 \rightarrow L (20.6)				
	2 ¹ A	2.45, 506	0.2834	H-1 \rightarrow L (78.3)	2.58, 480	0.0802		
Vitisin B	1 ¹ A	2.87, 432	0.0069	H \rightarrow L (19.9)	2.95, 420	0.0155	2.53, 491 ^e	
	2 ¹ A	2.49, 498	0.3616	H-2 \rightarrow L (98.0)	2.71, 458	0.5473		
				H-1 \rightarrow L (18.2)				
MAD ^f		0.08		H-1 \rightarrow L (81.8)	2.93, 423	0.0337		
				H \rightarrow L (17.5)				
				H-2 \rightarrow L (94.7)	3.22, 385	0.0072		
		0.06						

^a Experimental absorption maxima (eV, nm) are also given for comparison. ^b TD-PBE0/SV(P)//PBE0/SV(P) level of theory. ^c See ref 30. ^d See ref 31. ^e See ref 33. ^f Absolute mean deviation MAD (eV) for the maximum absorption band *in vacuo* and water (CPCM model).

Table 2. Calculated Excitation Energies ΔE (eV, nm), Main Configuration (Percentage Contribution in Parentheses), and Oscillator Strengths f (*in vacuo* and water), for Pigment A, Pinotin A, Mv-3-glc-vinylguaicol, and Pt-3-glc-4-vinylphenol^a

molecule	state	TDDFT ^b						exptl.
		vacuum			c-pcm (water)			
		ΔE (eV, nm)	f	configuration	ΔE (eV, nm)	f		
Pigment A	1 ¹ A	2.42, 513	0.7028	H \rightarrow L (97.5)	2.60, 477	0.7512	2.47, 503; ^{c, d} 2.46, 504 ^e	
	2 ¹ A	2.75, 452	0.0976	H-1 \rightarrow L (92.7)	2.98, 417	0.0385		
	3 ¹ A	3.07, 403	0.3578	H-1 \rightarrow L (92.7)	3.10, 400	0.2467		
Pinotin A	1 ¹ A	2.44, 509	0.8389	H-2 \rightarrow L (87.8)	2.63, 471	0.8232	2.42, 512; ^c 2.44, 509 ^e	
	2 ¹ A	2.73, 455	0.1262	H \rightarrow L (98.4)	2.85, 436	0.0722		
	3 ¹ A	2.84, 436	0.0291	H-1 \rightarrow L (87.3)	3.15, 393	0.0221		
Mv-3-glc- vinylguaicol	1 ¹ A	2.41, 514	0.8119	H-2 \rightarrow L (89.2)	2.58, 480	0.8358	2.42, 512 ^c	
	2 ¹ A	2.73, 455	0.1492	H-1 \rightarrow L (57.9)	2.87, 433	0.1184		
	3 ¹ A	2.82, 439	0.0636	H-2 \rightarrow L (40.4)	2.98, 416	0.0098		
Pt-glc-4-vinylphenol	1 ¹ A	2.82, 439	0.0636	H-2 \rightarrow L (57.4)	2.98, 416	0.0098	2.47, 503 ^c	
	2 ¹ A	2.82, 439	0.0636	H-1 \rightarrow L (38.9)	2.65, 467	0.7964		
	3 ¹ A	2.82, 439	0.0636	H \rightarrow L (95.4)	2.93, 423	0.0109		
		0.005		H-2 \rightarrow L (92.4)	3.11, 399	0.2396		

^a Experimental absorption maxima (eV, nm) are also given for comparison. ^b TD-PBE0/SV(P)//PBE0/SV(P) level of theory. ^c See ref 30. ^d See ref 31. ^e See ref 33. ^f Absolute mean deviation MAD (eV) for the maximum absorption band *in vacuo* and water (CPCM model).

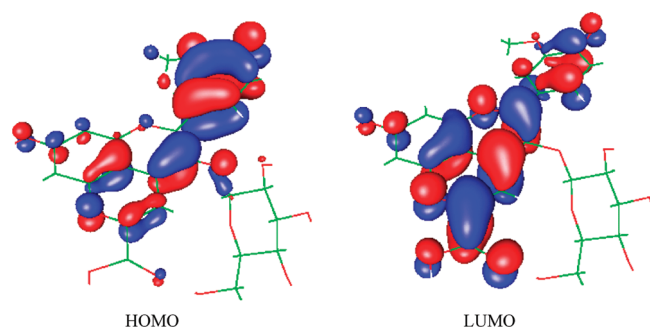


Figure 2. Isodensity molecular orbital plots (isodensity value of 0.03 au) of the HOMO and LUMO for the Pn-3-glc pyruvic acid derivative.

498 nm). For the other two molecules in Table 1 (Pn- and Pt-3-glc pyruvic adducts), the absorption maxima can be identified by their most intense transition (respectively at 539 and 527 nm). The inclusion of bulk solvation effects, through the CPCM model, lowers the absorption maxima and slightly intensifies the transition strengths, giving a better agreement with the experimental λ_{\max} . Moreover, for Vitisin-type molecules, the second excitation energies, as obtained from gas-phase calculations, become less intense ($f \sim 0.03$ – 0.08), and the main electronic band is mainly described by the first excitation energy. For *in vacuo* and water solution calculations, the absorption maxima roughly correlate with the HOMO–LUMO DFT energy gap ($E_{\text{H-L}}$). For example, for Vitisin-A, λ_{\max} is slightly greater (about 15 nm) than those of the Pn- and Pt-3-glc pyruvic derivatives, and its $E_{\text{H-L}}$ is decreased. In a similar way, Vitisin-B shows the lowest absorption maximum, among the compounds in Table 1, and the greatest $E_{\text{H-L}}$ value (in water 3.06 eV). Frontier molecular orbital energies (*in vacuo* and water) for all of the investigated compounds are reported in the Supporting Information (pp S17–S18).

The calculated λ_{\max} for pyruvic adducts does not strictly follow the experimental trend. The reason could be that the experimental values, as reported in Table 1, cover a small wavelength range (503–507 nm), while the TDDFT error, as found also for other classes of organic molecules, is generally within 0.3 eV or, in wavelength units, 1 or 2 orders of magnitude greater than the experimental difference.⁵⁶ The only case in which we can make a clear correlation between wavelength absorption maxima and molecular structure is Vitisin-B. This molecule, unlike pyruvic adducts, lacks the carboxylic group at position 12 of ring D that increases the electronic delocalization and consequently the molecule λ_{\max} . For Vitisin B, the calculated λ_{\max} decreases with respect to the other pyruvic acid adducts in agreement with the experimental trend. A rational explanation for the λ_{\max} bathochromic shift of pyruvic acid derivatives with respect to Vitisin B can be derived from the energy and wave function analysis of the frontier molecular orbitals. The decreasing of the HOMO–LUMO energy gap $\Delta E_{\text{H-L}}$ for Pn-3- and Pt-3-glc pyruvic acid adducts and Vitisin-A (gas-phase $\Delta E_{\text{H-L}} = 2.69$, 2.63, and 2.50 eV, see Table S1 in the Supporting Information) in comparison to that of Vitisin-B ($\Delta E_{\text{H-L}} = 2.79$ eV) is due to a greater LUMO energy stabilization. For example, the gas-phase HOMO's molecular energy difference between Pn-3-glc pyruvic acid and Vitisin-B is 0.43 eV (in water 0.22 eV) and lower than the LUMO's corresponding value of 0.53 eV (in water 0.39 eV). Similar results are found for Pt-3-glc pyruvic acid adducts and Vitisin-A in comparison to Vitisin-B with a better LUMO energy

stabilization. The molecular orbital composition analysis can help in the identification of which molecular structural factors cause the LUMO energy stabilization. For that purpose, the Mulliken electron population analysis procedure was followed for calculating the percentage weight of each atomic orbital to the HOMO and LUMO DFT orbitals.^{68,69} The molecular structures of Vitisin-B and pyruvic acid derivatives have been considered as composed by four main molecular fragments: the central molecular core (fused A, C, and D rings), ring B, glycosyl, and, in the case of pyruvic acid derivatives, carboxylic groups. In this way, the electron density is partitioned according to the molecular fragment weight composition. The central fused ring (A, C, and D) fragment accounts for about 20% of the total electron density in the HOMO wave function composition. The most important percentage contribution comes from ring B (77–80), while the glycosyl and the carboxylic groups make a negligible contribution. On the other hand, for the LUMO orbital composition, ring B contributes 15–21% and fused rings A, C, and D 15–17%, and an important contribution is derived from the carboxylic group (9%), which is absent in the case of the Vitisin-B molecule. These results can explain the important role of carboxylic substituent groups in the LUMO energy stabilization and consequently the wavelength bathochromic shift in pyruvic acid derivatives with respect to Vitisin-B.

The influence of the conformational equilibria on the UV–vis spectra has been analyzed for the case of Pn-3-glc pyruvic derivative in the gas phase. The rotation around the dihedral angle α (see Figure 1) for that compound can give rise to another minima conformer (see Supporting Information, p S14) with a small energy difference in comparison to the more stable structure, shown in Figure 1 ($\Delta E = 1.1$ kcal/mol). Neglecting in a first approximation the presence of other conformers, and applying the Boltzmann distribution formula at room temperature to the two conformers, we find that the resulting statistically weighted λ_{\max} over the relative populations is about 536 nm. This value is almost identical to the *in vacuo* value of 539 nm (see Table 1). In a water solution, the energetic difference between the two conformers is lower (0.02 kcal/mol). Also in this case, the weighted λ_{\max} (502 nm) is very close to the corresponding value calculated for the most stable conformer ($\lambda_{\max} = 506$ nm). For that reason, the calculation of the electronic spectrum of the most stable conformer has been assumed as a valid approximation for the spectra of the molecule weighted in all possible energy minima conformations. The mean absolute deviations (MAD) for the most intense transition *in vacuo* and in a water solution, as extrapolated from the spectra simulation, are respectively 0.08 and 0.06 eV, showing an overall good agreement between experimental and calculated excitation energy in the visible region. In particular, the inclusion of solvent effects on wavelength maxima improves the quality of the results. The simulated electronic spectra in a water medium, for all compounds in Table 1, are shown in Figure 3 (left part) along with the experimental spectrum of Vitisin A that consists of an intense band in the visible region and other bands in the ultraviolet (UV) part of the spectrum. The calculated line shape of the spectrum also reproduces qualitatively the minor UV peaks ($\lambda < 400$ nm).

For the vinyl derivatives reported in Table 2, the experimental wavelength maxima range between 503 nm (Pigment A and Pt-3-glc-vinylphenol) and 512 nm (Mv-3-glc-vinylguaicol), while *in vacuo* calculated values are between 509 (Pinotin A) and 520 nm (Pt-3-glc-vinylphenol). In solution, the theoretical λ_{\max} values are lowered and range between 471 and 486 nm. The MAD for this group of compounds, calculated for the *in vacuo* most intense

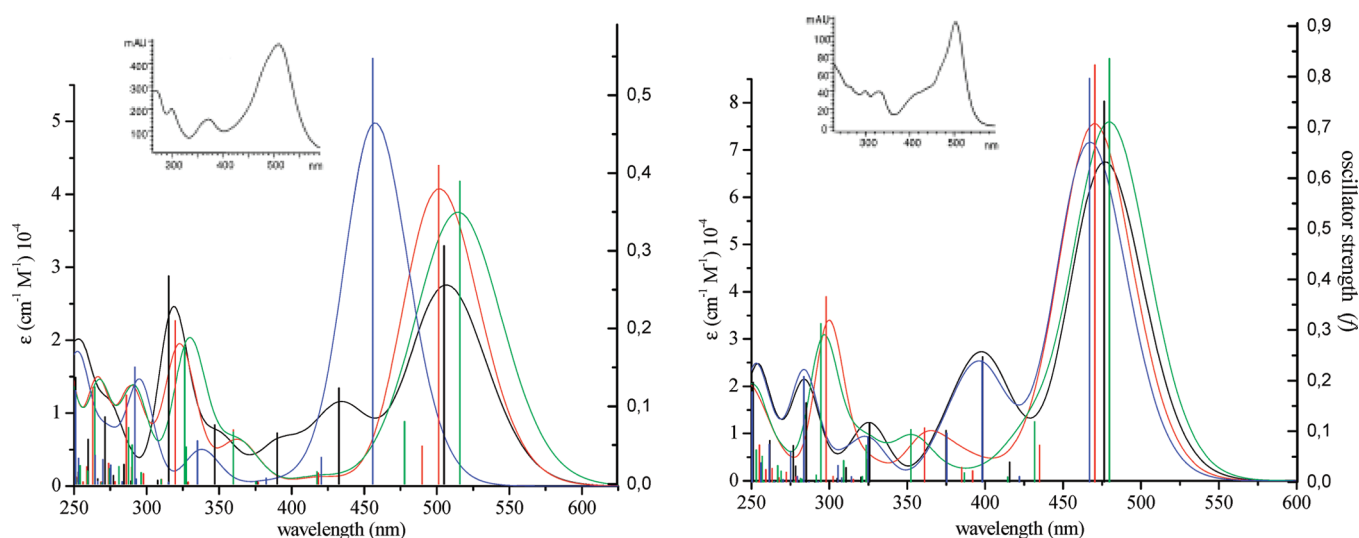


Figure 3. Simulated convoluted and stick electronic spectra in water of studied pyranoanthocyanins in Scheme 1 (absorption wavelength vs oscillator strength and molar absorptivity). On the left: Pn-3-glc pyruvic acid (black), Pt-3-glc pyruvic acid (red), Vitisin-A (green), and Vitisin-B (blue). On the right: Pigment A (black), Pinotin A (red), Mv-3-glc-vinylguaiacol (green), and Pt-3-glc-4-vinylphenol (blue). Experimental spectra of Vitisin-A (upper left part) and pigment A (upper right part) as taken from ref 30 are also reproduced.

Table 3. Calculated (RICC2) Excitation Energies ΔE (eV, nm), Main Configuration (Percentage Contribution in Parentheses), Singles Excitation Contribution T_1 (%), and Oscillator Strengths f for Peonidin- and Petunidin-3-Glucoside Pyruvic Acids (Pn-3-glc and Pt-3-glc der.), Vitisin A, and Vitisin B^a

molecule	state	RICC2 ^b					exptl.
		ΔE (eV, nm)	configuration	f	T_1		
Pn-3-glc der.	1 ¹ A	2.37, 524.1	H→L (93.8)	0.6704	85.7	2.47, 503 ^c	
	2 ¹ A	2.97, 417.6	H-1→L (65.9) H-2→L (28.1)	0.0946	86.6		
	3 ¹ A	3.34, 371.3	H-2→L (64.3) H-1→L (23.4)	0.0683	86.0		
Pt-3-glc der.	1 ¹ A	2.33, 533.1	H→L (93.4)	0.7526	86.2	2.45, 507 ^c	
	2 ¹ A	2.64, 469.4	H-1→L (91.9)	0.1970	85.7		
	3 ¹ A	3.05, 406.2	H-2→L (91.9)	0.0196	86.0		
Vitisin A	1 ¹ A	2.15, 575.6	H→L (92.4)	0.7029	84.7	2.43, 511 ^d	
	2 ¹ A	2.67, 464.2	H-1→L (89.8)	0.1174	85.5		
	3 ¹ A	3.10, 399.8	H-2→L (91.7)	0.0178	86.4		
Vitisin B	1 ¹ A	2.36, 524.6	H→L (94.1)	0.8054	85.4	2.53, 491 ^c	
	2 ¹ A	2.93, 423.2	H-1→L (90.1)	0.0602	85.9		
	3 ¹ A	3.26, 380.6	H-2→L (90.3)	0.0065	86.4		
MAD ^f						0.17	

^a Experimental absorption maxima (eV, nm) are also given for comparison. ^b RICC2/SV(P)//PBE(0)/SV(P) level of theory. ^c See ref 30. ^d See ref 31. ^e See ref 33. ^f Absolute mean deviation MAD (eV) for the *in vacuo* maximum absorption band.

excitation energy is 0.005 eV, whereas in a water solution the deviation is higher (0.17 eV). In Figure 2 (upper right part) is shown the experimental spectrum of Pigment A, which, apart from the main peak at 503 nm, has a broad minor peak centered at 423 nm. This electronic band corresponds to the third calculated excitation energy at 400 nm ($f = 0.2467$), which in the convoluted spectrum in Figure 2 is centered at 397 nm, showing good agreement with the experimental band. The final MAD for the eight pyranoanthocyanins is 0.04 eV *in vacuo* and 0.11 eV in water. The results obtained from *in vacuo* RICC2 calculations are reported in Tables 3 and 4. In addition to the TDDFT data types given in Tables 1 and 2 (e.g., excitation energies, configurations, and oscillator strengths), the percentage

contribution of singles substitution (T_1) is also indicated, which gives the quality of the RICC2 excitation energies, in comparison to a full configuration interaction treatment. In this case, T_1 values are all between 84% and 87%; this means that electronic transitions have a small double character contribution. For pyruvic and acetaldehyde derivatives (Table 3), the experimental wavelength maxima increasing order (from Pn-3-glc to Vitisin B) is qualitatively reproduced with an absolute mean deviation of 0.17 eV, which is greater than that found for the vinyl derivatives (0.07 eV). The MAD for all investigated systems as obtained from RICC2 gives a value of 0.12 eV that is comparable to the corresponding values for water solution TDDFT calculations.

Table 4. Calculated RICC2 Excitation Energies ΔE (eV, nm), Main Configuration (Percentage Contribution in Parentheses), Singles Excitation Contribution T_1 (%) and Oscillator Strengths f for Pigment A, Pinotin A, Mv-3-glc-vinylguaiacol, Pt-3-glc-4-vinylphenol^a

molecule	state	RICC2 ^b		f	T_1	exptl. ^c
		ΔE (eV, nm)	configuration			
Pigment A	1 ¹ A	2.33, 531.2	H→L (94.4)	1.0553	85.4	2.47, 503; ^{cd} 2.46, 504 ^e
	2 ¹ A	3.00, 414.0	H-1→L (55.9) H-2→L (29.6) H-3→L (7.4)	0.1775	85.8	
	3 ¹ A	3.16, 392.7	H-2→L (61.8) H-1→L (26.2) H-3→L (4.1)	0.2116	85.9	
Pinotin A	1 ¹ A	2.36, 525.2	H→L (94.0)	1.1884	85.6	2.42, 512; ^c 2.44, 509 ^e
	2 ¹ A	2.84, 436.9	H-1→L (92.8)	0.2552	85.3	
	3 ¹ A	3.13, 395.8	H-2→L (81.8)	0.0182	85.4	
Mv-3-Glu-4-vinylguaiacol	1 ¹ A	2.31, 535.7	H→L (94.3)	1.1449	85.4	2.42, 512 ^e
	2 ¹ A	2.86, 434.1	H-1→L (90.1)	0.2801	85.3	
	3 ¹ A	3.06, 405.2	H-2→L (83.8)	0.0277	85.3	
Pt-Glu-4-vinylphenol	1 ¹ A	2.50, 495.8	H→L (93.4)	1.1356	86.2	2.47, 503 ^c
	2 ¹ A	3.04, 408.3	H-1→L (68.8) H-2→L (19.0) H-3→L (5.1)	0.0720	85.7	
	3 ¹ A	3.17, 391.5	H-2→L (72.6) H-1→L (17.1)	0.2157	86.0	
MAD ^f						0.07

^a Experimental maxima wavelengths (eV, nm) are also given for comparison. ^b RICC2/SV(P)//PBE(0)/SV(P) level of theory. ^c See ref 30. ^d See ref 31. ^e See ref 33. ^f Absolute mean deviation MAD (eV) for the *in vacuo* maximum absorption band.

4. CONCLUSIONS

The theoretical electronic spectra of a group of anthocyanin-derived pigments have been computed by means of TDDFT and RICC2 methods. A preliminary conformational analysis was performed over all of the investigated pyranoanthocyanins in order to identify the most stable conformer within each. As for anthocyanins, the simultaneous rotation of ring C and substituents at position 12' of ring D (Figure 1) around carbon–carbon single bonds gives rise to several energy minima. Steric and intramolecular hydrogen bond factors can rationalize the relative energy stability of each conformer, whose difference is within 10 kcal/mol of the most stable conformer taken as an absolute energy minimum reference.

Calculated TDDFT excitation energies both *in vacuo* and in a water solution (CPCM model) showed mean absolute deviations of 0.04 and 0.11 eV, respectively. The latter value is comparable with that of the more refined RICC2 method, which gives MAD equal to 0.12 eV. Wavelength maxima absorption shifts due to molecular structural differences have been theoretically predicted between piruvic and acetaldehyde acid adducts. In particular, the presence of the electron-withdrawing carboxylic group attached on ring D for Pn-, Pt-, and Mv-3-glc derivatives causes, in accordance with the experimental findings, a bathochromic wavelength shift with respect to Vitisin B, due to the lowering of the HOMO–LUMO energy gap. However, the small differences between pyranoanthocyanin experimental wavelength maxima do not allow one to correctly predict, for all the sets of molecules at the TDDFT level, the λ_{\max} experimental changes according to the molecular structure differences. The Gaussian convolution of the excitation energies and oscillator strengths, for the solution simulation of the electronic spectra profile over the UV–visible part of the spectrum, gives a qualitatively good agreement with the experimental absorption line shape profiles as shown in Figure 3 for the cases of Vitisin-A and Mv-3-glv vinylguaiacol.

■ ASSOCIATED CONTENT

S Supporting Information. Optimized Cartesian coordinates in water for pyranoanthocyanins reported in Figure 1, optimized structures for other possible conformers with absolute and relative energies, frontier molecular orbital energies. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*Tel.: +39-0984-492048. Fax: +39-0984-492044. E-mail: nrusso@unical.it.

■ ACKNOWLEDGMENT

The University of Calabria, the Food Science and Engineering Interdepartmental Center of the University of Calabria, and L.I. P.A.C., Calabrian Laboratory of Food Process Engineering (Regione Calabria APQ - Ricerca Scientifica e Innovazione Tecnologica I atto integrativo, Azione 2 laboratori pubblici di ricerca mission oriented interfiliara, and Azione 3 sostegno alla domanda di innovazione nel settore agroalimentare) are gratefully acknowledged.

■ REFERENCES

- (1) Castañeda-Ovando, A.; Pacheco-Hernández, M.; Páez-Hernández, E.; Rodríguez, J. A.; Galán-Vidal, C. A. *Food Chem.* **2009**, *113*, 859–871.
- (2) Yoshida, K.; Mori, M.; Kondo, T. *Nat. Prod. Rep.* **2009**, *26*, 884–915.
- (3) Veitch, N. C.; Grayer, R. J. *Nat. Prod. Rep.* **2008**, *25*, 555–611.
- (4) Andersen, Ø. M.; Jordheim, M. In *Flavonoids: Chemistry, Biochemistry and Applications*; Andersen, Ø. M., Markham, K. R., Eds.; Taylor & Francis Group: Oxford, U. K., 2009; Chapter 10, pp 471–553.
- (5) Pauling, L. *Fortschr. Chem. Org. Nat.* **1939**, *3*, 203–235.

- (6) Hatier, J-H. B.; Gould, K. S. In *Anthocyanins: biosynthesis, functions, and applications*; Gould, K., Davies, K., Winefield, C., Eds.; Springer Science: New York, 2009; Chapter 1, pp 1–12.
- (7) Kähkönen, M. P.; Heinonen, M. J. *Agric. Food Chem.* **2003**, *51*, 628–633.
- (8) Halliwell, B. J. *Sci. Food Agric.* **2006**, *86*, 1992–1995.
- (9) Record, I. R.; Dreosti, I. E.; McInerney, J. K. *Br. J. Nutr.* **2001**, *85*, 459–464.
- (10) Fossen, T.; Cabrita, L.; Andersen, Ø. M. *Food Chem.* **1998**, *63*, 435–440.
- (11) Cabrita, L.; Fossen, T.; Andersen, Ø. M. *Food Chem.* **2000**, *68*, 101–107.
- (12) Wrolstad, R. E.; Durst, R. W.; Lee, J. *Trends Food Sci. Technol.* **2006**, *16*, 423–428.
- (13) Heredia, F. J.; Francia-Aricha, E. M.; Rivas-Gonzalo, J. C.; Vicario, I. M.; Santos-Buelgab, C. *Food Chem.* **1998**, *63*, 491–498.
- (14) Mateus, N.; de Freitas, V. In *Anthocyanins: biosynthesis, functions, and applications*; Gould, K., Davies, K., Winefield, C., Eds.; Springer Science: New York, 2009; Chapter 9, pp 283–298.
- (15) Borkowsky, T.; Szymusiak, H.; Gliszczynska-Swiglo, A.; Tyrakowska, B. *Food Res. Int.* **2005**, *38*, 1031–1037.
- (16) Fuyuki, I.; Tanaka, N.; Katsuki, A.; Fujii, T. *J. Photochem. Photobiol., A* **2002**, *150*, 153–157.
- (17) Fleschhut, J.; Kratzer, F.; Rechkemmer, G.; Kulling, S. E. *Eur. J. Nutr.* **2006**, *45*, 7–18.
- (18) Cameira dos Santos, P. J.; Brillouet, J. M.; Cheynier, V.; Moutounet, M. *J. Sci. Food Agric.* **1996**, *70*, 204–208.
- (19) Rentzsch, M.; Schwarz, M.; Winterhalter, P. *Trends Food Sci. Technol.* **2007**, *18*, 526–534.
- (20) Bakker, J.; Timberlake, C. F. *J. Agric. Food Chem.* **1997**, *45*, 35–43.
- (21) Fulcrand, H.; Benabdeljalil, C.; Rigaud, J.; Cheynier, V.; Moutounet, M. *Phytochemistry* **1998**, *47*, 1401–1407.
- (22) Romero, C.; Bakker, J. *J. Agric. Food Chem.* **1999**, *47*, 3130–3139.
- (23) Schwarz, M.; Wabnitz, T. C.; Winterhalter, P. *J. Agric. Food Chem.* **2003**, *51*, 3682–3687.
- (24) Schwarz, M.; Hofmann, G.; Winterhalter, P. *J. Agric. Food Chem.* **2004**, *52*, 498–504.
- (25) Schwarz, M.; Wray, V.; Winterhalter, P. *J. Agric. Food Chem.* **2004**, *52*, 5095–5101.
- (26) Hillebrand, S.; Schwarz, M.; Winterhalter, P. *J. Agric. Food Chem.* **2004**, *52*, 7331–7338.
- (27) Rein, M. J.; Ollilainen, V.; Vahermo, M.; Yli-Kauhaluoma, J.; Heinonen, M. *Eur. Food Res. Technol.* **2005**, *202*, 239–244.
- (28) Marston, A.; Hostettmann, K. In *Flavonoids: Chemistry, Biochemistry and Applications*; Andersen, Ø. M., Markham, K. R., Eds.; Taylor & Francis Group: Oxford, U. K., 2009; Chapter 1, pp 1–32.
- (29) Welch, C. R.; Wu, Q.; Simon, J. E. *Curr. Anal. Chem.* **2008**, *4*, 75–101.
- (30) Alcade-Eon, C.; Escribano-Baiolon, M. T.; Santos-Buelga, C.; Rivas-Gonzalo, J. C. *Anal. Chim. Acta* **2004**, *513*, 305–318.
- (31) De Villiers, A.; Vanhoenacker, G.; Majek, P.; Sandra, P. *J. Chromatogr., A* **2004**, *1054*, 195–204.
- (32) Alcade-Eon, C.; Escribano-Baiolon, M. T.; Santos-Buelga, C.; Rivas-Gonzalo, J. C. *Anal. Chim. Acta* **2004**, *563*, 238–254.
- (33) Vergara, C.; Mardones, C.; Hermosín-Gutiérrez, I.; von Baer, D. *J. Chromatogr., A* **2010**, *1217*, 5710–5717.
- (34) Giusti, M. M.; Wrolstad, R. E. In *Current Protocols in Food Analytical Chemistry*; Wrolstad, R. E., Ed.; John Wiley & Sons: New York, 2001; No. Unit F1.2.1-13.
- (35) Asenstorfer, R. E.; Iland, P. G.; Tate, M. T.; Jones, G. J. *Anal. Biochem.* **2003**, *318*, 291–299.
- (36) Torskangerpoll, K.; Børve, K. J.; Andersen, Ø. M.; Leif J. Sæthre, L. J. *Spectrochim. Acta, Part A* **1999**, *55*, 761–771.
- (37) Sakata, K.; Saito, N.; Honda, T. *Tetrahedron* **2006**, *62*, 3721–3731.
- (38) Woodford, J. N. *Chem. Phys. Lett.* **2005**, *410*, 182–187.
- (39) Ferreira da Silva, P.; Lima, J. C.; Freitas, A. A.; Shimizu, K.; Maçanita, A. L.; Quina, F. H. *J. Phys. Chem. A* **2005**, *109*, 7329–7338.
- (40) Leopoldini, M.; Rondinelli, F.; Russo, N.; Toscano, M. *J. Agric. Food Chem.* **2010**, *58*, 8862–8871.
- (41) Estévez, L.; Mosquera, R. A. *J. Phys. Chem. A* **2008**, *112*, 10614–10623.
- (42) Carvalho, A. R. F.; Oliveira, J.; de Freitas, V.; Silva, A.; Mateus, N.; Melo, A. *THEOCHEM* **2010**, *946*, 113–118.
- (43) Carvalho, A. R. F.; Oliveira, J.; de Freitas, V.; Mateus, N.; Melo, A. *THEOCHEM* **2010**, *948*, 61–64.
- (44) Freitas, A. A.; Shimizu, K.; Dias, L. G.; Quina, F. H. *J. Braz. Chem. Soc.* **2007**, *18*, 1537–1546.
- (45) Casida, M. E. *THEOCHEM* **2010**, *914*, 3–18.
- (46) Hättig, C.; Weigend, F. *J. Chem. Phys.* **2000**, *113*, 5154–5161.
- (47) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165–169.
- (48) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822–8824.
- (49) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.
- (50) Eichkorn, K.; Treutler, O.; Öhm, H.; Häser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *242*, 652–660.
- (51) Sierka, M.; Hogekamp, A.; Ahlrichs, R. *J. Chem. Phys.* **2003**, *118*, 9136–9148.
- (52) Schäfer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *97*, 2571–2577.
- (53) Eichkorn, K.; Weigend, F.; Treutler, O.; Ahlrichs, R. *Theor. Chem. Acc.* **1997**, *97*, 119–124.
- (54) Ernzerhof, M.; Scuseria, G. E. *J. Chem. Phys.* **1999**, *110*, 5029–5036.
- (55) Perdew, J. P.; Ernzerhof, M.; Burke, K. *J. Chem. Phys.* **1996**, *105*, 9982–9985.
- (56) Schäfer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829–5835.
- (57) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007–1023.
- (58) Quartarolo, A. D.; Sicilia, E.; Russo, N. *J. Chem. Theory Comput.* **2009**, *5*, 1849–1857.
- (59) Quartarolo, A. D.; Russo, N.; Sicilia, E.; Lelj, F. *J. Chem. Theory Comput.* **2007**, *3*, 860–869.
- (60) Jacquemin, D.; Perpète, E. A.; Ciofini, I.; Adamo, C. *Acc. Chem. Res.* **2009**, *42*, 326–334.
- (61) Perpète, E. A.; Jacquemin, D. *THEOCHEM* **2009**, *914*, 100–105.
- (62) Klamt, A.; Schuurmann, G. *J. Chem. Soc., Perkin Trans. 2* **1996**, *5*, 799–805.
- (63) Klamt, A.; Jonas, V. *J. Chem. Phys.* **1996**, *105*, 9972–9981.
- (64) Gorelsky, S. I. *SWizard Program*, revision 4.2; York University: Ontario, Canada, 1998. <http://www.sg-chem.net/swizard/> (accessed March 2011).
- (65) Hättig, C.; Hellweg, A.; Köhn, A. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1159–1169.
- (66) Cramer, J. C.; Truhlar, D. G. *J. Am. Chem. Soc.* **1993**, *115*, 5745–5153.
- (67) Molteni, C.; Parrinello, M. *J. Am. Chem. Soc.* **1998**, *120*, 2168–2171.
- (68) Mulliken, R. S. *J. Chem. Phys.* **1995**, *23*, 1833–1834.
- (69) Mulliken, R. S. *J. Chem. Phys.* **1995**, *23*, 1841–1846.

Assessment of TD-DFT and CC2 Methods for the Calculation of Resonance Raman Intensities: Application to *o*-Nitrophenol

Julien Guthmüller*

Institut für Physikalische Chemie, Friedrich Schiller Universität Jena, Helmholtzweg 4, 07743, Jena, Germany

ABSTRACT: The resonance Raman (RR) intensities of *o*-nitrophenol (oNP) were investigated theoretically with the aim of assessing the accuracy of excited state gradients calculated with DFT and CC2 approaches. It is found that the B3LYP and B2PLYP exchange-correlation (XC) functionals provide the best estimate of the ground state properties, while the other considered approaches present significantly less accurate vibrational frequencies and normal coordinates. Then, it is demonstrated that the use of the B3LYP force field for the ground state properties, in association with XC functionals including a large amount of HF exchange (M06-2X) or including long-range corrections (CAM-B3LYP and ω B97X) for the excited state gradient calculations, provides the most accurate RR spectra. Moreover, it is found that the RR intensities calculated with the best XC functionals show comparable accuracy to the results obtained with CC2 calculations. Finally, it is seen that the accuracy of the excited state gradients does not correlate with the accuracy of the excitation energies and oscillator strengths, for which XC functionals with a lesser amount of HF exchange (B3LYP, M06, and HSE06) provide more accurate results in the case of oNP. This indicates that the assessment of excited state gradients via the calculation of RR intensities, can provide additional information about the performance of quantum chemistry approaches in predicting excited state properties.

1. INTRODUCTION

The calculation of excited state properties for large molecules remains a challenge in quantum chemistry. Therefore, several studies have assessed the accuracy of time-dependent density functional theory (TDDFT) and of wave function based methods to calculate excitation energies of singlet and triplet excited states.^{1–6} The oscillator strengths associated with transitions between singlet states were also investigated in some works.^{7,8} However, much less is known about the accuracy of currently used quantum chemistry methods concerning the estimation of excited state gradients. Because excited state gradients are fundamentally important for a correct evaluation of excited state geometries and potential energy surfaces and for the subsequent treatment of excited state dynamics, a better knowledge of the performance of standard computational approaches is highly desirable.

An evaluation of the gradients can in principle be obtained from the calculated excited state geometries. However, this quantity is hard to evaluate, because experimental excited state geometries are usually not available. Nevertheless, a comparison can be made between experimental results and the calculated 0–0 excitations as well as with the simulated vibronic structure of the absorption spectrum, which can be obtained from a calculation of the Franck–Condon (FC) factors.^{9–12} For example, Dierksen and Grimme¹² found that the vibronic structure of a series of large compounds is better reproduced with exchange-correlation (XC) functionals incorporating about 30–40% Hartree–Fock (HF) exchange. However, such a comparison hardly allows an estimation of the calculated geometrical displacements along the individual coordinates, due to the usually limited resolution of the vibronic structure in the experimental spectra. On the other hand, resonance Raman (RR) spectroscopy^{13–15} provides the possibility to assess the excited state gradients

separately, i.e., along each vibrational coordinate. Indeed, within the so-called short-time approximation¹⁶ (STA), the RR intensities are directly related to the excited state gradients evaluated at the FC point. In this respect, the calculation of RR intensities and their comparison with experimental data offers an opportunity to gain more knowledge about the ability of standard quantum chemistry methods to determine excited state gradients.

This study initiates such an investigation by considering the prototype molecule of *o*-nitrophenol (oNP) (Figure 1). The RR spectrum of this compound was already investigated experimentally by Wang et al.¹⁷ in a cyclohexane solution, and it was shown to present a rich pattern of 15 RR active vibrations. The RR measurements were performed in resonance with the first absorption band, which is associated with the first singlet excited state. First, this experimental band displays a large broadening with no resolved vibronic structure. Second, the measured RR spectra at excitation wavelengths of 355 and 369 nm show small differences in their relative RR intensities. These facts indicate that the use of the STA for the evaluation of the relative RR intensities is adequate in this case. Thus, the purpose of this contribution is to assess the accuracy of several XC functionals within the framework of TDDFT calculations as well as of the second-order approximate coupled cluster singles and doubles¹⁸ (CC2) methods. These computational approaches are widely applied for determining excited state properties, mostly due to their good compromise between accuracy and computational cost. Therefore, an initial investigation of their performance is useful before a larger set of systems is considered.

The paper is organized as follows. Section 2 describes the employed approximations and computational methods. Section

Received: January 5, 2011

Published: March 24, 2011

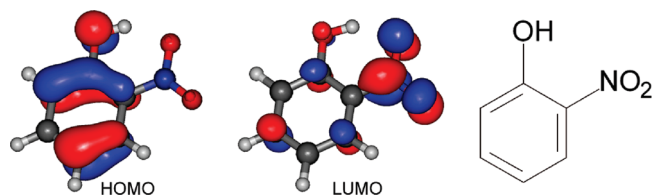


Figure 1. Structure of *o*-nitrophenol and molecular orbitals involved in the dominant excited state (B3LYP).

3.1 provides the assignment of the oNP vibrational frequencies, which is required before the RR spectra can be assessed. The vertical excitation energies and oscillator strengths are presented in section 3.2. Then, the accuracy of the considered theoretical methods for the evaluation of RR intensities is discussed in section 3.3, and conclusions are given in section 4.

2. COMPUTATIONAL METHODS

The geometry, harmonic vibrational frequencies, and normal coordinates of the ground state were obtained with the Gaussian 09 program¹⁹ by means of density functional theory (DFT) and second-order Møller–Plesset²⁰ (MP2) calculations. The DFT calculations were performed with the BLYP,^{21,22} B3LYP,^{23,22} HSE06²⁴ (HSEh1PBE), CAM-B3LYP,²⁵ M06-2X,²⁶ ω B97X,²⁷ and B2PLYP²⁸ XC functionals in association with the 6-311++G(2df,p) basis set. To correct for the lack of anharmonicity and the approximate treatment of electron correlation,²⁹ the harmonic frequencies obtained with the B3LYP, B2PLYP, HSE06, MP2, CAM-B3LYP, M06-2X, and ω B97X force fields were scaled by factors of 0.98, 0.98, 0.96, 0.96, 0.95, 0.95, and 0.95, respectively. Additionally, the effects of the solvent cyclohexane ($\epsilon = 2.0165$) on the ground state properties were taken into account by the integral equation formalism of the polarizable continuum model³⁰ (IEFPCM).

The vertical excitation energies, oscillator strengths, and analytical Cartesian energy derivatives of the excited states (gradients) were obtained from TDDFT calculations employing the 6-311++G(2df,p) basis set. The TDDFT calculations were performed by using the same XC functionals as for the ground state properties. Moreover, the excited state properties were also determined with all previously mentioned XC functionals in addition to M06,²⁶ B3LYP-35,³¹ BMK,³² and ω B97²⁷ by employing the B3LYP and B2PLYP ground state geometries, frequencies, and normal coordinates. The effects of the solvent were approximated with the IEFPCM model, and the nonequilibrium procedure of solvation was used for the computation of the excitation energies and excited state gradients.

The excitation energies, oscillator strengths, and gradients of the excited states were also evaluated with the second-order approximate coupled cluster singles and doubles¹⁸ (CC2) and spin-component scaled CC2⁵ (SCS-CC2) approaches. These calculations were performed with the RICC2 module^{33,34} of the TURBOMOLE 6.2 program,³⁵ thus making use of the resolution of the identity approximation. The def2-QZVPP basis set³⁶ and its associated auxiliary basis set were employed for all (SCS)-CC2 calculations. Moreover, SCS-CC2 computations made use of scale parameters⁵ for the opposite-spin and same-spin components equal to $c_{os} = 6/5$ and $c_{ss} = 1/3$, respectively. All (SCS)-CC2 calculations were performed in a vacuum using the B3LYP, B2PLYP, and MP2 ground state geometries obtained with the Gaussian 09 program.

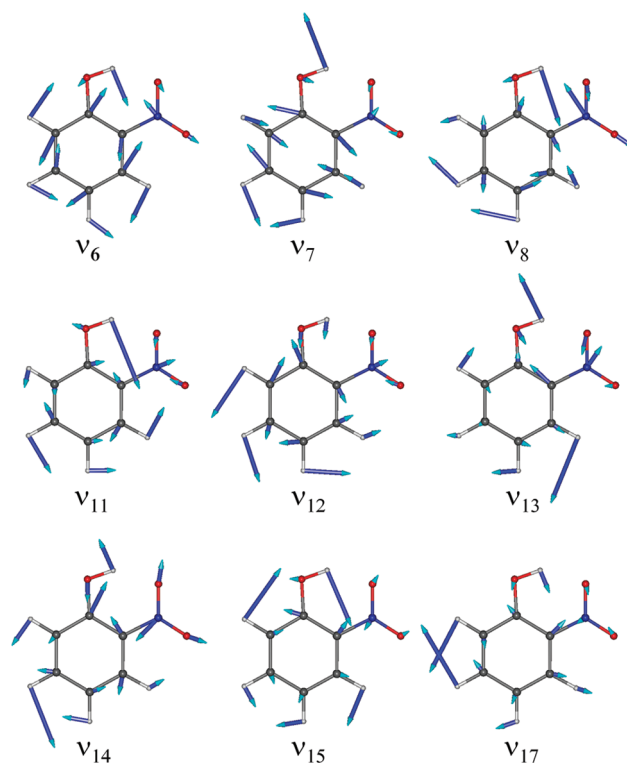


Figure 2. Nuclear displacements of several vibrational normal modes of *o*-nitrophenol calculated at the B3LYP/6-311++G(2df,p) level of approximation.

The relative RR intensities were obtained within the short-time approximation.¹⁶ In the STA, the RR intensity for a fundamental transition $0 \rightarrow 1_l$ can be obtained from the partial derivatives of the excited state electronic energy (E^e) along the l th normal coordinate (Q_l) evaluated at the ground state equilibrium geometry:

$$I_{0 \rightarrow 1_l} \propto \frac{1}{\omega_l} \left(\frac{\partial E^e}{\partial Q_l} \right)_0^2 \quad (1)$$

where ω_l is the frequency of the l th normal mode. These gradients were obtained from the analytical derivatives of the excited state electronic energy (E^e) along the nonmass-weighted Cartesian coordinates according to the relation

$$\left(\frac{\partial E^e}{\partial Q} \right)_0 = L^T M^{-1/2} \left(\frac{\partial E^e}{\partial x} \right)_0 \quad (2)$$

where M is the matrix containing the atomic masses, L is the orthogonal matrix obtained from the solution of the ground state normal mode eigenvalue problem and connects the mass-weighted Cartesian coordinates to the mass-weighted normal coordinates, and $(\partial E^e / \partial Q)_0$ and $(\partial E^e / \partial x)_0$ are column vectors containing the derivatives along the normal coordinates and Cartesian coordinates, respectively.

Finally, the relative nonresonant Raman intensities at the standard excitation wavelength of 1064 nm were obtained according to the relation

$$I_{0 \rightarrow 1_l} \propto \omega_L (\omega_L - \omega_l)^3 (45a_l^2 + 7\gamma_l^2) \quad (3)$$

where a_l^2 and γ_l^2 are invariants for randomly oriented molecules^{37,15} evaluated from the analytical derivatives of the

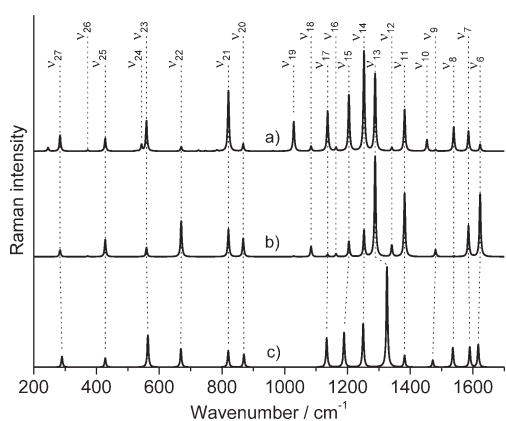


Figure 3. Assignment of the vibrational frequencies. (a) B3LYP/6-311++G(2df,p) Raman spectrum in cyclohexane at 1064 nm. (b) B3LYP/6-311++G(2df,p) RR spectrum in cyclohexane calculated within the STA. (c) Experimental RR spectrum in cyclohexane at 355 nm; the spectrum is reconstructed from the cross-sections reported in Table 4 of ref 17. All of the spectra are normalized with respect to the most intense band, and a Lorentzian function with a fwhm of 5 cm^{-1} is employed to broaden the transitions.

polarizability tensor along the normal coordinates $(\partial\alpha_{ij}/\partial Q_i)_0$. These derivatives were obtained from the B3LYP ground state vibrational frequency calculation performed with Gaussian 09.

3. RESULTS AND DISCUSSION

3.1. Assignment of the *o*-Nitrophenol Vibrational Frequencies. To assess the accuracy of the calculated RR intensities and therefore the excited state gradients, it is first necessary to assign unambiguously the experimental frequencies to the theoretical vibrational modes (Figure 2). Assignments of the oNP vibrations were already reported in the literature^{38,17} and were deduced by comparing theoretical (DFT/B3LYP) frequencies and IR intensities with experimental frequencies obtained from IR, Raman, and RR spectra. Therefore, the assignment of the fundamental transitions in the $200\text{--}1700 \text{ cm}^{-1}$ wavenumber range is first re-examined. This is performed by simulating the Raman and RR intensities and by comparing them to the previously reported experimental spectra. To this aim, Figure 3 shows the calculated Raman and RR spectra obtained with the B3LYP XC functional as well as the experimental RR spectrum, which was reconstructed from the RR cross-sections reported in Table 4 of ref 17. It should be mentioned that the same mode numbering as the one employed by Kovács et al.³⁸ and Wang et al.¹⁷ is used for the fundamental vibrations with an in-plane symmetry. Additionally, the calculated vibrational frequencies are reported in Table 1 and are compared to experimental RR frequencies recorded in cyclohexane solution and to Raman frequencies recorded in CCl_4 solution.

The assignment deduced by comparing the theoretical (B3LYP) and experimental Raman spectra is found in excellent agreement with the one reported by Kovács et al.³⁸ The only exception concerns the experimental Raman active vibration at 668 cm^{-1} , which is here assigned to the ν_{22} mode at 670 cm^{-1} (B3LYP) instead of a vibration of out-of-plane symmetry. It can also be mentioned that the ν_9 , ν_{12} , and ν_{26} vibrations were not assigned, because they are not visible in the experimental Raman spectrum. This fact is in agreement with their weak calculated

Table 1. Experimental and Calculated Vibrational Frequencies

no.	calculated frequencies ^a (cm^{-1})		experimental frequencies (cm^{-1})	
	B3LYP	B2PLYP	RR ^b	Raman ^c
ν_6	1623.0	1627.2	1617	1623
ν_7	1585.9	1593.0	1590	1593
ν_8	1538.8	1539.1	1538	1540
ν_9	1480.8	1482.9	1472	
ν_{10}	1453.4	1458.6		1459
ν_{11}	1382.2	1387.9	1382	1382
ν_{12}	1341.3	1347.4		
ν_{13}	1288.2	1292.9	1326	1327
ν_{14}	1252.7	1257.4	1250	1255
ν_{15}	1204.5	1210.4	1190	1192
ν_{16}	1163.4	1164.0		1159
ν_{17}	1136.9	1139.9	1134	1139
ν_{18}	1084.2	1084.1		1081
ν_{19}	1028.9	1031.5		1030
ν_{20}	867.7	863.9	870	871
ν_{21}	820.5	816.3	820	820
ν_{22}	670.0	667.5	669	668
ν_{23}	559.5	558.7	564	564
ν_{24}	543.7	540.6		547
ν_{25}	427.9	424.0	428	426
ν_{26}	372.2	371.2		
ν_{27}	283.8	284.1	290	286
MAD ^d	6.6	8.9		
MAX ^d	37.8 (ν_{13})	33.1 (ν_{13})		
MAD ^e	5.4	5.8		
MAX ^e	38.8 (ν_{13})	34.1 (ν_{13})		

^aThe calculated harmonic frequencies are scaled by a factor of 0.98.

^bExperimental RR frequencies in cyclohexane solution.¹⁷ ^cExperimental Raman frequencies in CCl_4 solution.³⁸ ^dMean absolute deviation (MAD) and maximal absolute deviation (MAX) with respect to the experimental RR frequencies. ^eMAD and MAX with respect to the experimental Raman frequencies.

Raman intensity (Figure 3). However, these three vibrations (ν_9 , ν_{12} , ν_{26}) with frequencies of 1480.8 , 1341.3 , and 372.2 cm^{-1} , respectively, can be assigned to the experimental IR bands in CCl_4 solution³⁸ at 1479 , 1333 , and 372 cm^{-1} , respectively.

The comparison between the theoretical and experimental RR spectra provides an assignment in global agreement with the one reported by Wang et al.¹⁷ The only exception concerns the most intense RR band in the experimental spectrum at 1326 cm^{-1} , which is assigned here to the vibration ν_{13} at 1288.2 cm^{-1} instead of the vibration ν_{12} at 1341.3 cm^{-1} . This assignment is strongly motivated by the larger RR intensity calculated for ν_{13} in comparison to ν_{12} and is also in agreement with the assignment of the Raman spectrum. However, it should be noted that the frequency of ν_{13} shows the maximal absolute deviation (MAX) with respect to experimental results with an underestimation of about 38 cm^{-1} . This deviation of the ν_{13} frequency was already noted by Kovács et al.³⁸ and was possibly attributed to a Fermi resonance interaction. Such effects are not included in the present calculations, which employ the harmonic approximation. Moreover, the experimental IR frequencies in a CCl_4 solution³⁸ of the ν_{12} and ν_{13} vibrations are close, with values of 1333 and

Table 2. Vertical Excitation Energies (E^e) and Oscillator Strengths (f) of the First Singlet Excited State

	B3LYP geometry		B2PLYP geometry	
	E^e (eV) ^a	f	E^e (eV) ^a	f
BLYP	2.94 (−0.63)	0.0477	2.91 (−0.66)	0.0458
ω B97 ^b	4.18 (0.61)	0.1749	4.18 (0.61)	0.1722
SCS-CC2 ^c	4.08 (0.51)	0.1035	4.08 (0.51)	0.1016
ω B97X	4.07 (0.50)	0.1619	4.07 (0.50)	0.1588
M06-2X	4.00 (0.43)	0.1376	3.99 (0.42)	0.1342
CC2 ^c	3.97 (0.40)	0.1065	3.97 (0.40)	0.1046
BMK	3.86 (0.29)	0.1188	3.85 (0.28)	0.1152
CAM-B3LYP	3.84 (0.27)	0.1302	3.82 (0.25)	0.1266
B3LYP	3.38 (−0.19)	0.0763	3.37 (−0.20)	0.0734
B3LYP-35	3.71 (0.14)	0.1032	3.69 (0.12)	0.0996
M06	3.45 (−0.12)	0.0848	3.43 (−0.14)	0.0816
HSE06	3.50 (−0.07)	0.0835	3.48 (−0.09)	0.0804
exptl. ^d	3.57	0.0689	3.57	0.0689

^a The energy deviations with respect to experimental results are given in brackets. ^b Energies and oscillator strengths of the second singlet excited state. ^c Energies and oscillator strengths calculated in a vacuum. ^d Experimental energy and oscillator strength in a cyclohexane solution.¹⁷

1325 cm^{−1}, respectively. Therefore, it might be possible that the RR intensities of these two bands are superimposed in the experiment.

The vibrational frequencies of oNP were also calculated with the double-hybrid B2PLYP XC functional. It was shown recently³⁹ that this functional provides reliable vibrational frequencies and normal coordinates. From Table 1 it is seen that B2PLYP gives vibrational frequencies in close agreement with the B3LYP results. The larger deviation is found for the ν_7 vibration with a difference of about 7 cm^{−1}. It can also be mentioned that the normal coordinates obtained with both theoretical methods show similar nuclear displacements. Additionally, the mean absolute deviations (MAD) of the vibrational frequencies with respect to experimental results are rather small for both XC functionals, with MADs between 5.4 and 8.9 cm^{−1}. Such values are in agreement with previous works^{29,39–42} and confirm the reliability of the B3LYP and B2PLYP functionals for the determination of ground state vibrational frequencies. Therefore, the geometries, frequencies, and normal coordinates obtained with these two force fields will be employed in the following to investigate the RR intensities of oNP.

3.2. Excited States. The vertical excitation energy and oscillator strength of the first allowed excited state of oNP were calculated with different XC functionals as well as with (SCS)-CC2 methods by employing the B3LYP and B2PLYP ground state geometries (Table 2). The experimental absorption spectrum in cyclohexane¹⁷ shows a single unstructured band in the 400–300 nm range, which can be associated with the first singlet excited state of oNP. For each considered method, this state involves a transition from the HOMO to the LUMO orbitals (Figure 1), which present a transfer of electronic density going from the aromatic cycle and OH group to the NO₂ group. Therefore, a geometrical reorganization is expected to occur over the entire molecule after excitation to the excited state. This is also in agreement with the fact that several fundamental vibrations show RR activity in the 200–1700 cm^{−1} wavenumber

range and are associated with normal coordinates distributed on the different parts of oNP.

From Table 2, it is seen that small differences are found between the excitation energies (<0.03 eV) and oscillator strengths calculated with B3LYP and B2PLYP. Of course, this is related to the similar geometries obtained with both methods, which present bond length differences lower than 0.004 Å. The comparison between the different XC functionals shows that the excitation energies and oscillator strengths are strongly dependent on the amount of HF exchange included in the functional. Indeed, the excitation energy calculated with the pure GGA functional BLYP is underestimated by 0.63 eV in comparison to experimental results (B3LYP geometry), whereas the hybrid functionals B3LYP, M06, and HSE06 including a moderate amount of HF exchange of 20, 27, and 25%, respectively, improve significantly the value of the excitation energy. Thus, the best agreement with respect to experimental results is found for the screened hybrid functional HSE06 with an underestimation of the excitation energy of only 0.07 eV (B3LYP geometry). Then, functionals with a larger amount of HF exchange show an overestimation of the excitation energy, as can be seen for B3LYP-35, BMK, and M06-2X, which incorporate 35, 42, and 54% of HF exchange, respectively. Of these functionals, B3LYP-35 provides the most accurate results with an overestimation of only 0.14 eV. Moreover, rather significant overestimations are obtained with the long-range corrected functionals CAM-B3LYP, ω B97, and ω B97X as well as with the (SCS)-CC2 methods, with CAM-B3LYP providing the best estimate. However, it should be mentioned that no solvent effects were included in the (SCS)-CC2 calculations. By approximating the solvatochromic shift with a B3LYP/IEFPCM calculation, the (SCS)-CC2 energies should be decreased by about 0.1 eV, which improves the CC2 excitation energy to a value very close to the BMK and CAM-B3LYP results. Moreover, it is seen in Table 2 that the oscillator strengths are overestimated in comparison to experimental results for all methods, except BLYP. Similarly to the excitation energies, the most accurate oscillator strengths are obtained with the XC functionals incorporating a moderate amount of HF exchange, i.e., B3LYP, HSE06, and M06, whereas the other methods provide oscillator strengths that are significantly overestimated. In the next step, how the accuracy on the excitation energies and oscillator strengths correlates to the accuracy on the RR spectra will be investigated.

3.3. Effect of the Computational Method on the o-Nitrophenol RR Spectrum. **3.3.1. RR Spectra Calculated with the B3LYP Force Field.** The STA RR spectra calculated with the B3LYP force field for the ground state and with different theoretical methods for the evaluation of the excited state gradients are presented in Figure 4 and are compared to the experimental spectrum. In order to provide a more quantitative comparison between the theoretical methods, the MAD and MAX of the relative RR intensities with respect to experimental results are reported in Table 3. First, it is seen that BLYP gives the larger MAD and shows a RR spectrum with too strong intensities in the 1300–1700 cm^{−1} wavenumber range and an incorrect intensity pattern for the ν_{14} , ν_{15} , and ν_{17} vibrations. Improvements are obtained for most of the intensities with the four hybrid functionals M06, B3LYP, HSE06, and B3LYP-35. However, the vibrations ν_6 , ν_{11} , and ν_{22} still show noticeable overestimated intensities, whereas vibrations ν_8 and ν_{17} have a too weak RR intensity in comparison to experimental results. Despite

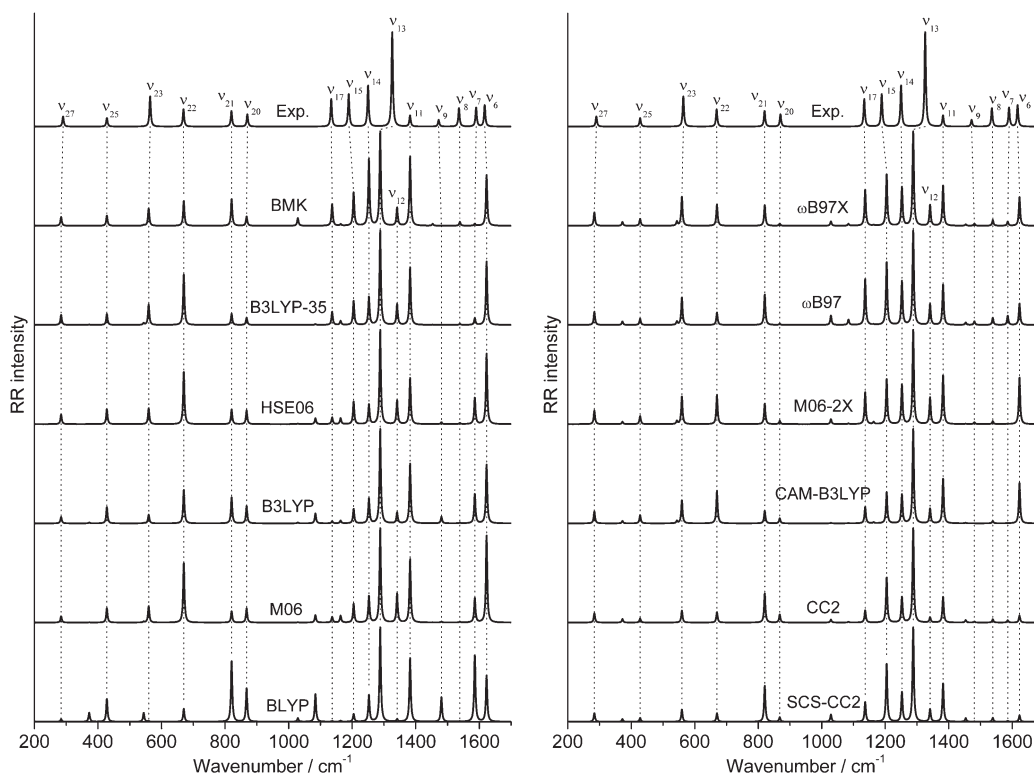


Figure 4. Comparison between the experimental RR spectrum¹⁷ and the STA RR spectra calculated using the B3LYP force field in association with different methods for the evaluation of excited state gradients. The spectra are normalized with respect to vibration ν_{13} , and a Lorentzian function with a fwhm of 5 cm^{-1} is employed to broaden the transitions.

a large MAX value for vibration ν_{11} , the BMK functional provides a further improved MAD, mostly due to a better description of the ν_6 , ν_{15} , ν_{17} , and ν_{22} RR intensities. Next, the RR spectra obtained with the (SCS)-CC2 methods present interesting differences with respect to the spectra calculated with DFT methods: (i) The vibrations ν_6 , ν_7 , and ν_8 are obtained with comparable intensities, which is in better agreement with experimental results even if their intensities are globally underestimated. (ii) Both ν_{11} and ν_{12} intensities are significantly reduced, which provides an additional confirmation for the assignment of vibrations ν_{11} , ν_{12} , and ν_{13} . (iii) CC2 gives the smallest value of MAX, with a deviation of only 0.193 for vibration ν_{23} , illustrating the overall good agreement obtained with this method. (iv) Similarly to the excitation energies, the use of SCS-CC2 does not provide an improvement in comparison to CC2. Moreover, smaller MADs are found with M06-2X and with the long-range corrected functionals CAM-B3LYP, ω B97, and ω B97X: (i) The ω B97X functional provides the smallest MAD with a value of only 0.094. (ii) The RR intensities of the low-frequency modes in the $200\text{--}900 \text{ cm}^{-1}$ wavenumber range are better reproduced with these functionals, even if the intensity of ν_{20} is underestimated with the ω B97(X) methods. (iii) The intensities of ν_{14} , ν_{15} , and ν_{17} are in overall good agreement with experimental results, despite an overestimation in the case of ω B97. (iv) the overestimation of the ν_{11} intensity is reduced in comparison to other functionals but is still larger than the one obtained with CC2. (v) The ω B97(X) functionals provide an improved description of the intensities in the $1400\text{--}1700 \text{ cm}^{-1}$ wavenumber range in comparison to the CAM-B3LYP, M06-2X, and (SCS)-CC2 methods. Finally, the comparison between the calculations and experimental results shows that the accuracy

Table 3. Mean Absolute Deviations (MAD) and Maximal Absolute Deviations (MAX) of the Relative RR Intensities with Respect to Experimental Results^a

	B3LYP force field		B2PLYP force field	
	MAD ^b	MAX	MAD	MAX
BLYP	0.264 (0.275)	0.559 (ν_{11})	0.257	0.631 (ν_{11})
M06	0.203 (0.222)	0.697 (ν_6)	0.231	0.875 (ν_{11})
B3LYP	0.180 (0.195)	0.516 (ν_{11})	0.192	0.686 (ν_{11})
HSE06	0.170 (0.189)	0.523 (ν_6)	0.192	0.623 (ν_{11})
B3LYP-35	0.162 (0.172)	0.493 (ν_{11})	0.187	0.727 (ν_{11})
BMK	0.148 (0.148)	0.622 (ν_{11})	0.190	0.782 (ν_{11})
SCS-CC2 ^c	0.141 (0.144)	0.290 (ν_{11})	0.133	0.404 (ν_{11})
CC2 ^c	0.120 (0.122)	0.193 (ν_{23})	0.118	0.216 (ν_{11})
CAM-B3LYP	0.116 (0.126)	0.356 (ν_{11})	0.133	0.511 (ν_{11})
M06-2X	0.115 (0.118)	0.400 (ν_{11})	0.134	0.621 (ν_{11})
ω B97	0.109 (0.108)	0.322 (ν_{15})	0.110	0.451 (ν_{11})
ω B97X	0.094 (0.097)	0.309 (ν_{11})	0.094	0.462 (ν_{11})

^a The experimental RR spectrum in cyclohexane solution at 355 nm is from ref 17. The comparison is made by normalizing to the intensity of vibration ν_{13} in both calculated and experimental RR spectra. ^b The MADs with respect to the experimental RR spectrum at 369 nm are given in brackets. ^c Excited state derivatives calculated in a vacuum.

of the simulated RR spectra (Table 3) is not correlated with the accuracy of the excitation energies and oscillator strengths (Table 2). This is clearly seen for the functionals ω B97(X), which provide the RR spectra with the lowest MADs but give the

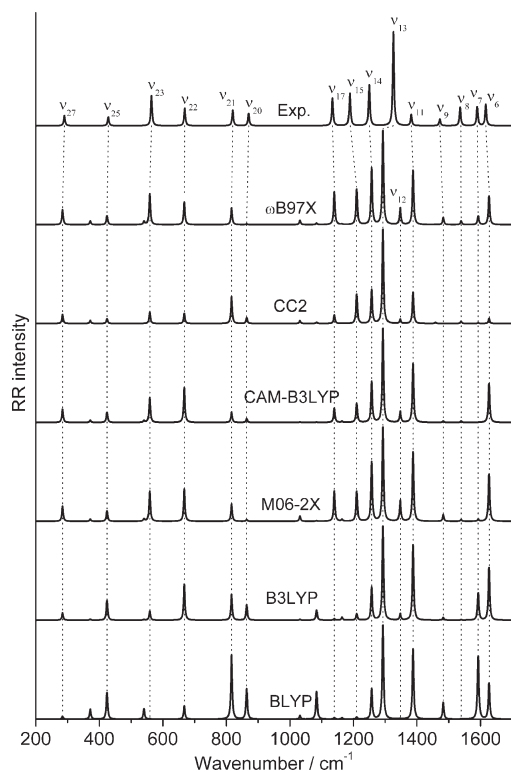


Figure 5. Comparison between the experimental RR spectrum¹⁷ and the STA RR spectra calculated using the B2PLYP force field in association with different methods for the evaluation of excited state gradients. The spectra are normalized with respect to vibration ν_{13} , and a Lorentzian function with a fwhm of 5 cm^{-1} is employed to broaden the transitions.

largest overestimations of both the excitation energy and the oscillator strength.

Additionally, in order to confirm the validity of the STA, the RR spectrum was simulated using the ω B97X functional by including the vibronic structure of the excited state according to a method described elsewhere.³¹ This allows the dependency of the RR spectrum with respect to the excitation wavelength to be accounted for. Thus, the simulation of the RR spectrum for an excitation wavelength of 355 nm provides a MAD of 0.102 and a MAX of 0.297 (for vibration ν_{11}). These values are very close to those obtained within the STA and consequently justify the use of this approximation for the purpose of assessing the accuracy of different theoretical methods. This is also corroborated by the small dependency of the experimental RR intensities with respect to the excitation wavelength.¹⁷ Indeed, the MADs obtained by comparing the simulated intensities with the experimental spectrum recorded for an excitation wavelength of 369 nm (Table 3) show the same trends as those obtained by comparing to the 355 nm experimental spectrum.

3.3.2. RR Spectra Calculated with the B2PLYP and Other Force Fields. The RR spectra obtained with the B2PLYP force field (Figure 5) present no significant improvements in comparison to those obtained with the B3LYP force field (Figure 4). Indeed, as can be seen from Table 3, most of the MADs are larger or very close to the one calculated with the B3LYP force field. The only small improvements are found for the MADs of the (SCS)-CC2 methods, which are closer to those obtained with the XC functionals CAM-B3LYP, M06-2X, and ω B97(X). Additionally,

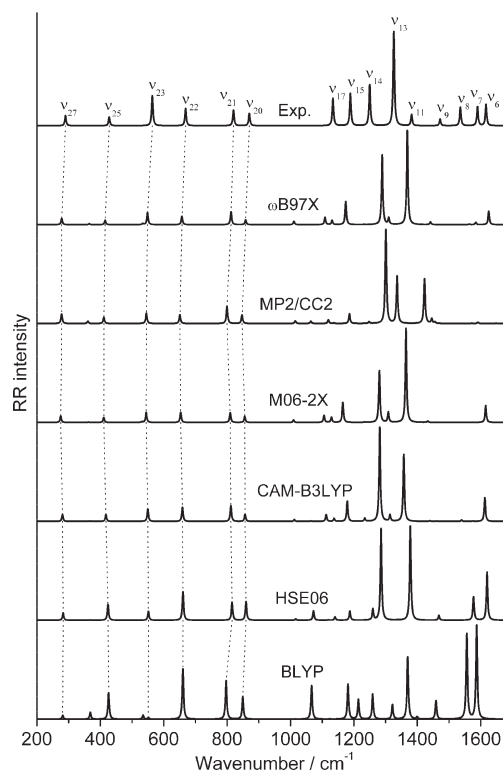


Figure 6. Comparison between the experimental RR spectrum¹⁷ and the STA RR spectra calculated using different computational methods. The spectra are normalized with respect to vibration ν_{13} , and a Lorentzian function with a fwhm of 5 cm^{-1} is employed to broaden the transitions.

all of the MAX values corresponding to vibration ν_{11} are increased in comparison to B3LYP, indicating that this mode is less accurately described with the B2PLYP force field. Therefore, on the basis of the vibrational frequencies (Table 1) and of the RR intensities (Table 3), it appears that B2PLYP does not improve the ground state properties of oNP in comparison to B3LYP.

Furthermore, it is interesting to investigate the accuracy of the RR spectra in the situation where the same theoretical method is employed for both the ground and the excited state calculations. Thus, Figure 6 shows the RR spectra obtained with the DFT methods BLYP, HSE06, CAM-B3LYP, M06-2X, and ω B97X as well as the spectrum obtained using MP2 for the ground state properties and CC2 for the excited state gradients. It should be mentioned that comparable normal coordinates are found with all methods for the ν_{20} , ν_{21} , ν_{22} , ν_{23} , ν_{25} , and ν_{27} vibrations, which leads to a straightforward assignment of these vibrations. However, the normal coordinates of all of the vibrations in the $1000\text{--}1700 \text{ cm}^{-1}$ wavenumber range are found to be strongly different between the theoretical methods. As can be seen from Figure 6, this leads to important differences in the simulated RR intensities in this wavenumber window. It appears also clearly that no spectrum improves over those calculated with the B3LYP or B2PLYP force fields, but instead large discrepancies with respect to experimental results are obtained. This also shows that an assignment of the vibrational bands in the $1000\text{--}1700 \text{ cm}^{-1}$ wavenumber range is hardly possible by employing these methods for the ground state properties of oNP and consequently that the geometries and normal coordinates obtained with these force fields are less accurate than those calculated with B3LYP and

B2PLYP. This result is also in agreement with a recent work³⁹ which showed that HSE06, CAM-B3LYP, M06-2X, and ω B97X provide less accurate vibrational frequencies than B3LYP or B2PLYP.

4. CONCLUSIONS

The RR relative intensities of *o*-nitrophenol were investigated theoretically with the aim of assessing the accuracy of excited state calculations based on DFT and CC2 approaches. The comparison between simulated Raman and RR spectra with experimental results allowed a reliable assignment of the vibrational bands. Thus, it is found that B3LYP provides the best estimate of the ground state properties, while B2PLYP calculations show no improvement and even a slightly reduced accuracy in comparison to B3LYP. Moreover, the results obtained with the BLYP, HSE06, CAM-B3LYP, M06-2X, ω B97X, and MP2 methods present significantly less accurate vibrational frequencies and normal coordinates, leading to important differences in their respective RR spectrum, which show large discrepancies with respect to experimental results. However, the use of the B3LYP force field for the ground state in association with different methods for the excited state gradients shows a noticeable improvement of the accuracy of RR intensities. Thus, XC functionals including a large amount of HF exchange and long-range corrections like M06-2X, CAM-B3LYP, and ω B97(X) provide the most accurate RR spectra. The RR intensities obtained with the best XC functionals are of comparable accuracy to those obtained with CC2 calculations, which shows that these approaches should be considered more often in the future for simulating RR intensities of middle-sized to large systems. It is also seen that the accuracy of the excited state gradients does not correlate with the accuracy of the excitation energies and oscillator strengths, for which XC functionals with lesser HF exchange like B3LYP, M06, and HSE06 provide more accurate results in the case of oNP. This result also indicates that in addition to evaluating vertical excitation energies, the assessment of excited state gradients offers the possibility of having a complementary view about the performance of newly developed functionals in predicting excited state properties. Finally, the study demonstrates that an accurate description of both ground and excited state properties remains a challenging task, even for recently developed theoretical approaches. However, a significant improvement of the RR relative intensities can be obtained using a hybrid approach, in which the ground and excited states are described by two separate theoretical methods.

AUTHOR INFORMATION

Corresponding Author

*E-mail: julien.guthmuller@uni-jena.de.

ACKNOWLEDGMENT

The author thanks the Carl-Zeiss Stiftung for financial support and the Thüringer Ministerium für Bildung, Wissenschaft und Kultur (PhotoMIC). All of the calculations have been performed at the Universitätsrechenzentrum of the Friedrich-Schiller University of Jena and on the HP computers of the Theoretical Chemistry group. The author also thanks Prof. Leticia González for helpful discussions as well as Martin Elstner and Marcus Schulze for performing preliminary calculations on oNP.

REFERENCES

- (1) Silva-Junior, M. R.; Schreiber, M.; Sauer, S. P. A.; Thiel, W. *J. Chem. Phys.* **2008**, *129*, 104103.
- (2) Jacquemin, D.; Wathelet, V.; Perpète, E. A.; Adamo, C. *J. Chem. Theory Comput.* **2009**, *5*, 2420–2435.
- (3) Jacquemin, D.; Perpète, E. A.; Ciofini, I.; Adamo, C. *J. Chem. Theory Comput.* **2010**, *6*, 1532–1537.
- (4) Goerigk, L.; Grimme, S. *J. Chem. Phys.* **2010**, *132*, 184103.
- (5) Hellweg, A.; Grün, S. A.; Hättig, C. *Phys. Chem. Chem. Phys.* **2008**, *10*, 4119.
- (6) Silva-Junior, M. R.; Schreiber, M.; Sauer, S. P. A.; Thiel, W. *J. Chem. Phys.* **2010**, *133*, 174318.
- (7) Schreiber, M.; Silva-Junior, M. R.; Sauer, S. P. A.; Thiel, W. *J. Chem. Phys.* **2008**, *128*, 134110.
- (8) Miura, M.; Aoki, Y.; Champagne, B. *J. Chem. Phys.* **2007**, *127*, 084103.
- (9) Guthmuller, J.; Zutterman, F.; Champagne, B. *J. Chem. Theory Comput.* **2008**, *4*, 2094–2100.
- (10) Guillaume, M.; Liégeois, V.; Champagne, B.; Zutterman, F. *Chem. Phys. Lett.* **2007**, *446*, 165–169.
- (11) Santoro, F.; Improta, R.; Lami, A.; Bloino, J.; Barone, V. *J. Chem. Phys.* **2007**, *126*, 084509.
- (12) Dierksen, M.; Grimme, S. *J. Phys. Chem. A* **2004**, *108*, 10225–10237.
- (13) Albrecht, A. C. *J. Chem. Phys.* **1961**, *34*, 1476.
- (14) Myers, A. B. *Chem. Rev.* **1996**, *96*, 911–926.
- (15) Long, D. A. *The Raman Effect: A Unified Treatment of the Theory of Raman Scattering by Molecules*; John Wiley & Sons Ltd: New York, 2002.
- (16) Heller, E. J.; Sundberg, R.; Tannor, D. *J. Phys. Chem.* **1982**, *86*, 1822–1833.
- (17) Wang, Y.; Wang, H.; Zhang, S.; Pei, K.; Zheng, X.; Lee Phillips, D. *J. Chem. Phys.* **2006**, *125*, 214506.
- (18) Christiansen, O.; Koch, H.; Jørgensen, P. *Chem. Phys. Lett.* **1995**, *243*, 409–418.
- (19) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, Revision A.02; Gaussian Inc.: Wallingford, CT, 2009.
- (20) Head-Gordon, M.; Head-Gordon, T. *Chem. Phys. Lett.* **1994**, *220*, 122–128.
- (21) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.
- (22) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
- (23) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
- (24) Henderson, T. M.; Izmaylov, A. F.; Scalmani, G.; Scuseria, G. E. *J. Chem. Phys.* **2009**, *131*, 044108.
- (25) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51–57.
- (26) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2007**, *120*, 215–241.
- (27) Chai, J.; Head-Gordon, M. *J. Chem. Phys.* **2008**, *128*, 084106.
- (28) Grimme, S. *J. Chem. Phys.* **2006**, *124*, 034108.
- (29) Merrick, J. P.; Moran, D.; Radom, L. *J. Phys. Chem. A* **2007**, *111*, 11683–11700.
- (30) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999–3094.


- (31) Guthmuller, J.; Champagne, B. *J. Chem. Phys.* **2007**, *127*, 164507.
- (32) Boese, A. D.; Martin, J. M. L. *J. Chem. Phys.* **2004**, *121*, 3405.
- (33) Hättig, C.; Weigend, F. *J. Chem. Phys.* **2000**, *113*, 5154.
- (34) Köhn, A.; Hättig, C. *J. Chem. Phys.* **2003**, *119*, 5021.
- (35) *TURBOMOLE V6.2*; University of Karlsruhe and Forschungszentrum Karlsruhe GmbH: Karlsruhe, Germany, 2010. Available from <http://www.turbomole.com> (accessed March 2011).
- (36) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297.
- (37) Wilson, E. B., Jr.; Decius, J. C.; Cross, P. C. *Molecular Vibrations*; McGraw-Hill: New York, 1955.
- (38) Kovács, A.; Izvekov, V.; Keresztury, G.; Pongor, G. *Chem. Phys.* **1998**, *238*, 231–243.
- (39) Biczysko, M.; Panek, P.; Scalmani, G.; Bloino, J.; Barone, V. *J. Chem. Theory Comput.* **2010**, *6*, 2115–2125.
- (40) Jiménez-Hoyos, C. A.; Janesko, B. G.; Scuseria, G. E. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6621.
- (41) Guthmuller, J.; González, L. *Phys. Chem. Chem. Phys.* **2010**, *12*, 14812.
- (42) Guthmuller, J.; Cecchet, F.; Lis, D.; Caudano, Y.; Mani, A. A.; Thiry, P. A.; Peremans, A.; Champagne, B. *ChemPhysChem* **2009**, *10*, 2132–2142.

Normal Mode Analysis in Zeolites: Toward an Efficient Calculation of Adsorption Entropies

Bart A. De Moor,[†] An Ghysels,[‡] Marie-Françoise Reyniers,^{*,†} Veronique Van Speybroeck,[‡] Michel Waroquier,[‡] and Guy B. Marin[†]

[†]Laboratory for Chemical Technology, Ghent University, Krijgslaan 281 S5, 9000 Ghent, Belgium

[‡]Center for Molecular Modeling, Ghent University, Technologiepark 903, 9052 Zwijnaarde, Belgium

 Supporting Information

ABSTRACT: An efficient procedure for normal-mode analysis of extended systems, such as zeolites, is developed and illustrated for the physisorption and chemisorption of *n*-octane and isobutene in H-ZSM-22 and H-FAU using periodic DFT calculations employing the Vienna Ab Initio Simulation Package. Physisorption and chemisorption entropies resulting from partial Hessian vibrational analysis (PHVA) differ at most $10 \text{ J mol}^{-1} \text{ K}^{-1}$ from those resulting from full Hessian vibrational analysis, even for PHVA schemes in which only a very limited number of atoms are considered free. To acquire a well-conditioned Hessian, much tighter optimization criteria than commonly used for electronic energy calculations in zeolites are required, i.e., at least an energy cutoff of 400 eV, maximum force of 0.02 eV/\AA , and self-consistent field loop convergence criteria of 10^{-8} eV . For loosely bonded complexes the mobile adsorbate method is applied, in which frequency contributions originating from translational or rotational motions of the adsorbate are removed from the total partition function and replaced by free translational and/or rotational contributions. The frequencies corresponding with these translational and rotational modes can be selected unambiguously based on a mobile block Hessian–PHVA calculation, allowing the prediction of physisorption entropies within an accuracy of $10\text{--}15 \text{ J mol}^{-1} \text{ K}^{-1}$ as compared to experimental values. The approach presented in this study is useful for studies on other extended catalytic systems.

1. INTRODUCTION

Molecular simulations are a valuable tool to obtain a better understanding of hydrocarbon adsorption and conversion processes in zeolites.^{1–21} Nowadays, theoretical calculations on extended zeolite systems are frequently performed as evidenced from the various studies in literature applying quantum mechanical/molecular mechanical (QM/MM),^{3–10} QM/QM^{11–15} and periodic density functional theory (DFT)^{16–21} methods. However, most studies reported in literature only concern electronic energies, while vibrational analysis is usually omitted since it is computationally very demanding. Nonetheless, frequency analysis is part and parcel of the study of hydrocarbon conversion reactions since they are required to calculate the rate and equilibrium coefficients that govern hydrocarbon conversions in practically relevant conditions of temperature and pressure.^{22–24} Consequently, criteria for geometry optimization and electronic energy calculation are well established, but the appropriate program settings related to the calculation of harmonic frequencies are less well documented. Nevertheless, a full Hessian vibrational analysis (FHVA) is more and more applied, e.g., in hybrid QM/MM methods.^{5,8–10} The combination of a high-level method for a small part of the system and a lower level method for the remainder of the zeolite structure makes a vibrational analysis feasible at a reasonable computational cost. In their hybrid QM/QM approach using the ONIOM method, McCann et al.,¹³ Lesthaeghe et al.,¹⁴ and Vandichel et al.¹⁵ applied a partial Hessian vibrational analysis (PHVA)^{25–31} on a 46T cluster model of H-ZSM-5, keeping the saturating hydrogen atoms

fixed. In contrast, for periodic DFT methods, very few studies report the extremely time-consuming vibrational analysis. The Hessian, being the second derivatives matrix of the energy with regard to atom displacements, is typically determined via numerical differentiation of the gradient by displacing the atoms, and its calculation thus leads to higher computational costs when the number of atoms in the zeolite unit cell increases. Svelle et al.¹² have studied methylation reactions of ethene, propene, and 2-butene in H-ZSM-5 (289 atoms/unit cell) and calculated the Hessian for a subset of atoms including the hydrocarbon atoms and the 56 zeolite atoms surrounding the acid site. They indicated that this PHVA reduces computational cost by 75% as compared to a FHVA. Tuma and Sauer²⁰ have performed a FHVA to study the (thermodynamic) stability of various isobutene complexes in H-FER (217 atoms/unit cell).

In this work, we present a periodic DFT study on the physisorption of *n*-octane and isobutene and the chemisorption of *t*-butyl carbenium ion, *i*-butoxy, and *t*-butoxy in H-ZSM-22 and H-FAU; these complexes are representative for the various types of physisorbed and chemisorbed complexes that occur in acid zeolite catalyzed alkane or alkene conversion processes. Figure 1 shows the structures of the various adsorption complexes in H-FAU and H-ZSM-22. We especially focus on the development of a cost-effective procedure for normal-mode analysis (NMA) that yields an acceptable accuracy for the

Received: September 24, 2010

Published: March 15, 2011

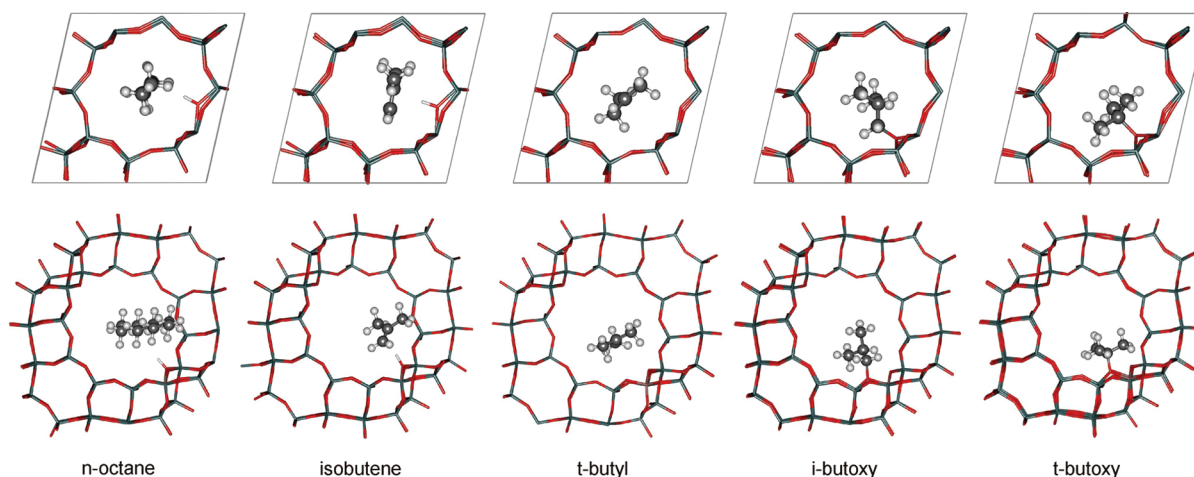


Figure 1. Physorption and/or chemisorption complexes of n-octane and isobutene in H-ZSM-22 and H-FAU (top and bottom figures, respectively).

calculation of enthalpies, entropies, and the equilibrium coefficients. For example, an accuracy of a factor of 2–6 on the adsorption equilibrium coefficient requires an accuracy of 6–15 J mol⁻¹ K⁻¹ on the adsorption entropy. To develop a cost-effective procedure, physisorption and chemisorption entropies resulting from FHVA are compared with results obtained using various PHVA schemes (see Section 2), considering a different number of “free” (N_{free}) or “fixed” (N_{fixed}) atoms for the numerical Hessian calculation. Obviously, a FHVA corresponds to a situation in which all atoms are free, and the gradient is calculated for $6N$ displaced structures, N being the total number of atoms in the unit cell. In case some atoms are kept fixed, the construction of the partial Hessian involves calculation of the gradient of $6N_{\text{free}}$ displaced structures. It is clear that by decreasing the number of free atoms, the vibrational analysis can be speeded up, as also indicated by Svelle et al.¹² The ultimate aim is to select a PHVA scheme with N_{free} as small as possible while approaching the accuracy of a FHVA. Physisorption and chemisorption entropies are the ideal quantities to evaluate the accuracy of a PHVA, since they are derived directly from the calculated harmonic frequencies using textbook statistical thermodynamics.³² The fact that entropies are extremely sensitive to even small variations in low-frequency values (0–100 cm⁻¹) makes them even more suited to evaluate the accuracy of PHVA schemes, as these low frequencies are omnipresent in the studied zeolite systems. On the other hand, physisorption and chemisorption enthalpies are much less sensitive to variations in harmonic frequency values and will not be discussed in detail. Previous work has shown that, in the temperature range from 300 to 800 K, the difference between physisorption energies and enthalpies of n-alkanes and alkenes amounts to some 0–5 kJ/mol at most.^{9,10}

From the systems studied in this work, physisorbed n-octane and isobutene and chemisorbed *t*-butyl carbenium ion are loosely bonded complexes in H-ZSM-22 and H-FAU zeolites: No chemical bond between the complex and the zeolite is present. In contrast, a C–O bond is formed in the *i*-butoxy and *t*-butoxy alkoxides after protonation of a physisorbed isobutene.

The absence of a direct bond between the adsorbed complex and the zeolite implies a higher translational and rotational mobility of the loosely bonded structures inside the zeolite pores. Previous work has shown that for n-alkanes physisorption

entropy losses calculated from harmonic frequencies only, i.e., the immobile adsorbate method, are overestimated as compared to experimental values.^{9,10} Therefore, a mobile adsorbate method has been proposed, in which some of the motions are not treated in the harmonic limit but rather considered to be free translational or rotational motions of the adsorbate in the zeolite pore. In the mobile adsorbate method,^{9,10} the partition function is calculated according to eq 1, in which $q_{\text{immobile}}^{\text{vibr}}$ is the total vibrational partition function (as generally obtained from standard simulation packages), from which the vibrational contribution corresponding to the n rotational and translational modes is removed (q_{nD}^{vibr}) and replaced by a n -dimensional (nD) free translational and rotational contribution ($q_{nD}^{\text{trans/rot}}$):

$$q_{\text{mobile}} = \frac{q_{\text{immobile}}^{\text{vibr}}}{q_{nD}^{\text{vibr}}} \times q_{nD}^{\text{trans/rot}} \quad (1)$$

The selection of frequencies corresponding to translational and/or rotational modes is however somewhat ambiguous as it depends on the user’s interpretation. In this work, the use of a mobile block Hessian (MBH) based approach to identify these frequencies is explored.

All FHVA calculations have been performed using the Vienna Ab Initio Simulation Package (VASP),^{33–36} and postprocessing for PHVA and MBH calculations have been done using the in-house developed software module TAMkin,^{37,38} a free python programmed versatile toolkit for normal-mode analysis and chemical kinetics.^{31,39–43}

2. METHODOLOGY

Figure 2 gives a schematic overview of all types of normal-mode analysis that have been performed in this paper. The following paragraphs comprise a discussion of these different types of methods, i.e., FHVA versus PHVA, the assumption of a mobile versus immobile adsorbate, the use of MBH for an unambiguous selection of translational and rotational modes, and finally the different PHVA schemes.

Full Hessian versus Partial Hessian Vibrational Analysis. It is well-known that an FHVA is one of the most time-consuming steps when studying extended systems, such as zeolites. Especially when the Hessian is calculated by numerical differentiation—as is the case for most periodic DFT simulation packages—the

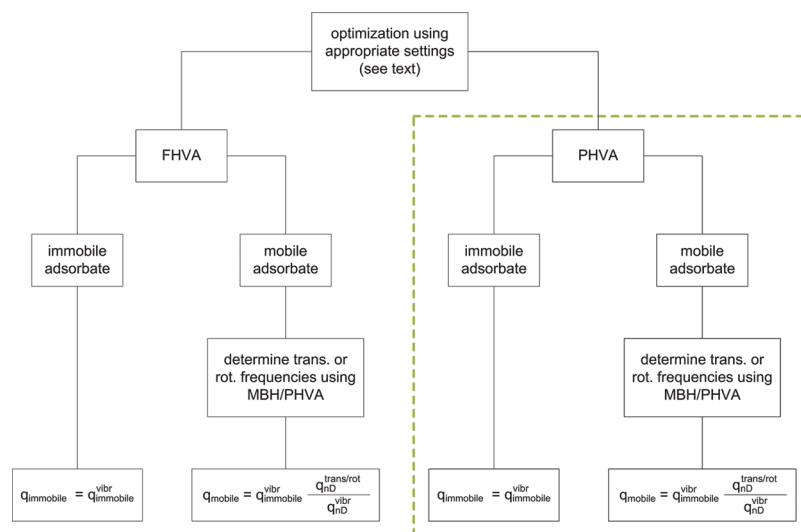


Figure 2. Overview of the types of normal-mode analysis that have been performed for the considered physisorption and chemisorption complexes in H-ZSM-22 and H-FAU.

computational effort is significant, due to the large number of atoms in the zeolite unit cell and the calculation of six displacements for each atom. PHVA has the advantage over FHVA that the derivatives with regard to the fixed atoms do not need to be calculated such that in total $6N_{\text{free}}$ gradient evaluations are required instead of $6N$. Many PHVA schemes can be proposed. The different schemes are characterized by the number of fixed atoms and by the particular selection of the fixed atoms. The effect of these parameters on the final outcome is not known for zeolite systems. Not only for small organic systems, such as alcohols and alkanes, but also for large macromolecules, such as proteins, this has been explored before.^{31,40–43}

All FHVA and PHVA calculations are performed with TAMkin.^{37,38} The TAMkin package loads the second derivatives matrix (Hessian) from the VASP output files. Mass-weighting and diagonalization of the full $3N \times 3N$ Hessian yields $3N$ frequencies and normal modes. TAMkin calculates $3N_{\text{free}}$ PHVA frequencies by first omitting the rows and columns corresponding to fixed atom displacements and next by diagonalizing this smaller Hessian of dimension $3N_{\text{free}} \times 3N_{\text{free}}$. No additional Hessian calculations were required to generate all PHVA results in this paper, as all PHVA Hessians could instantaneously be derived from the same FHVA Hessian by TAMkin.^{37,38}

Immobile versus Mobile Adsorbate. As mentioned in the Introduction, previous work has shown that, for loosely bonded physisorbed complexes, entropy losses calculated based on harmonic frequencies only, i.e., the immobile adsorbate method, are overestimated as compared to experimental values.^{9,10} The assumption of an immobile adsorbate conflicts with the significant translational and rotational degrees of freedom of the loosely bonded complex. Therefore a correction scheme, called the mobile adsorbate method, has been proposed (eq 1). The mobile adsorbate method was successfully applied in a study on physisorption of alkanes and alkenes in H-FAU, H-BEA, H-MOR, and H-ZSM-5.^{9,10} In eq 1, the partition function of the mobile adsorbate method, $q_{\text{immobiler}}$, is calculated from the total partition function of the immobile adsorbate method, $q_{\text{immobiler}}^{\text{vibr}}$, in which the contribution q_{nD}^{vibr} of the n harmonic frequencies (typically 3 or 4), corresponding with a translation and/or rotation of the hydrocarbon in the zeolite, is replaced by

the contribution $q_{nD}^{\text{trans/rot}}$ of n free translational and rotational contributions.

The n degrees of freedom in the mobile adsorbate method are determined from visualization of the FHVA normal modes. More specifically, translations and/or rotations of an adsorbate are considered to be mobile if the following criteria are fulfilled: (1) the harmonic frequencies lie in the range $0–100 \text{ cm}^{-1}$ and (2) the visualization of the corresponding normal modes shows a translation or rotation of the adsorbate in the zeolite as a whole with only small coupling to internal vibrations of the hydrocarbon or the zeolite. Based on these criteria, we observed the following number of mobile degrees of freedom (n): 2D free translation and 1D free rotation ($n = 3$) is assumed for *n*-octane physisorption in H-ZSM-22 and for isobutene physisorption and *t*-butyl chemisorption in both H-ZSM-22 and H-FAU zeolites, while 2D free translation and 2D free rotation ($n = 4$) is assumed for *n*-octane physisorption in H-FAU. The evaluation of the translational partition function requires a “molecular surface area” (analogously to the molecular volume in case of 3D free translation for gas-phase hydrocarbons). This has been chosen as $200 \times 600 \text{ pm}$ and $800 \times 800 \text{ pm}$ for H-ZSM-22 and H-FAU, respectively, according to the available zeolite pore dimensions. Figure 3 illustrates the 2D free translation and the 1D free rotation of the loosely bonded complexes in H-ZSM-22.

Until now, selection of the translational and rotational frequencies is done manually which is somewhat ambiguous and user dependent due to mixing of translational and rotational modes in the low-frequency range. In addition, this work is labor intensive in the case of FHVA calculations in view of the large number of frequencies in the $0–100 \text{ cm}^{-1}$ range. The MBH approach, in which the hydrocarbon is defined as a rigid block, offers an attractive alternative to this ad-hoc procedure in defining unambiguously these translational and rotational frequencies, as explained in the following section.

Mobile Block Hessian to Select Translational and Rotational Modes. In the MBH method, part of the system is grouped into rigid but mobile blocks. During the vibrational analysis, the blocks are allowed to translate or rotate as a whole, such that each block has 6 degrees of freedom (3 translational, 3 rotational), which can be described by 6 so-called block

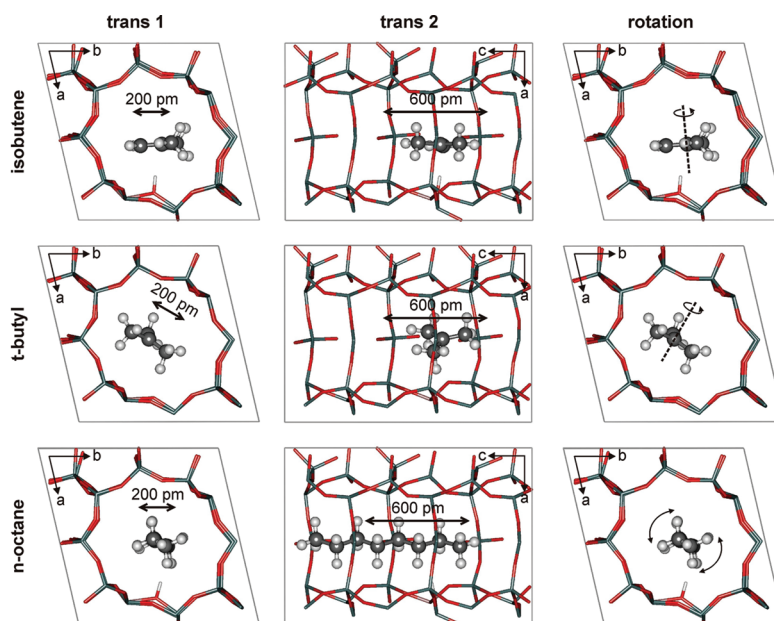


Figure 3. Impression of the 2D free translation and the 1D free rotation of the isobutene, *t*-butyl, and *n*-octane in H-ZSM-22.

parameters. The internal geometry of the block is kept fixed, which means internal vibrations within the block are not allowed. The number of degrees of freedom of the molecular system is thus reduced from $3N$ to $d = 6N_b + N_{free}$, where N_b is the number of blocks and N_{free} is the number of free atoms which are not part of any block. Whereas the $3N$ FHVA frequencies are calculated by mass weighting and diagonalizing the $3N \times 3N$ full Hessian, the introduction of the block concept requires the diagonalization of the smaller $d \times d$ MBH. This MBH Hessian consists of the second derivatives matrix of the potential energy with respect to the block parameters and to the coordinates of the free atoms. This MBH matrix can either be calculated by projection of the full Hessian or, if it would be implemented in the simulation package, be directly constructed by well chosen numerical differentiation. After appropriate mass weighting, its diagonalization yields the d MBH frequencies.^{39–43} In this paper, TAMkin is used to calculate the MBH frequencies.^{37,38} By making a smart block choice, one can focus on the interesting part of the frequency spectrum.

The combination of the mobile block concept of MBH with fixing zeolite atoms, as in PHVA, allows a more systematic approach to determine the values of n harmonic frequencies corresponding to the mobile translational and rotational motions of the mobile adsorbate method (see eq 1). In the MBH–PHVA approach in the present study, the zeolite atoms (except acid proton) are kept fixed at their reference position (PHVA concept), while the adsorbate molecule is a mobile block during the vibrational analysis (MBH concept). Whereas initially the n harmonic frequencies had to be selected out of the $3N$ FHVA modes, the MBH–PHVA approach reduces this to selecting n frequencies out of nine frequencies in case of isobutene and *n*-octane physisorption or out of six frequencies in case of chemisorption of the *t*-butyl carbenium ion. Indeed, by fixing all zeolite atoms except the acid proton and considering the adsorbate as a rigid mobile block, only nine frequencies are found from the combined MBH–PHVA approach in the case of physisorption of *n*-octane and isobutene. Three of them correspond to the vibration of the acid proton, and the six other frequencies

mainly correspond to translation and rotation of the hydrocarbon in the zeolite. In case of chemisorption of the *t*-butyl carbenium ion, the deprotonated zeolite is completely fixed, while the carbenium ion is defined as block, resulting in only six frequencies all corresponding to translation or rotation. This significantly simplifies the tedious identification of the n harmonic frequencies (for the calculation of q_{iD}^{vibr}), that are to be replaced by n free translational and rotational motions ($q_{immobile}^{vibr}$) in the mobile adsorbate method (eq 1). As a consequence, not only labor intensive visualization efforts are reduced but also the unambiguous identification of the n harmonic frequencies now becomes possible, since coupling with internal zeolite or hydrocarbon vibrations is excluded.

FHVA and Considered PHVA Schemes. A schematic overview of the FHVA and the various PHVA schemes considered in this work is shown in Figure 4. Case A represents the FHVA calculation: all atoms are considered free for the vibrational analysis in the gas-phase hydrocarbon, the unloaded zeolite and the zeolite–adsorbate system, i.e., six displacements for each atom are considered for the numerical Hessian calculation. Case B groups all PHVA calculations, with the number of free atoms (N_{free}) ranging from a few (B.[N_{free} low]) to almost all zeolite atoms free (B.[N_{free} high]). In cases A and B, hydrocarbon atoms are always considered free. Comparison of cases A and B allows the evaluation of the accuracy of the various PHVA schemes for the calculation of the physisorption and chemisorption entropies for the studied systems. Cases C and D are analogous to case A and B, but the hydrocarbon is defined as a mobile block without internal degrees of freedom. Comparison of the cases C and D with their analogues in cases A and B allows the assessment of the contribution of the internal hydrocarbon vibrations to the physisorption and chemisorption entropies in H-ZSM-22 and H-FAU.

Many different PHVA schemes have been included in this study. Table 1 summarizes all different choices, mentioning the name, the number of free and fixed zeolite atoms and the number of free or fixed T (Al or Si) atoms is mentioned between brackets. Starting from one free zeolite atom, i.e., the acid proton, we gradually increased the number of free Si or O atoms around the

Scheme	Type of Calculation	hydrocarbon	zeolite	hydrocarbon + zeolite
A	FHVA			
		FHVA	FHVA	FHVA
B.[N _{free} high]	PHVA			
		FHVA	PHVA	PHVA
⋮	⋮	⋮	⋮	⋮
B.[i]	PHVA
⋮	⋮	⋮	⋮	⋮
B.[N _{free} low]	PHVA			
		FHVA	PHVA	PHVA
C	MBH-FHVA			
		MBH	FHVA	MBH-FHVA
D.[N _{free} high] ^{MBH}	MBH-PHVA			
		MBH	PHVA	MBH-PHVA
⋮	⋮	⋮	⋮	⋮
D.[i] ^{MBH}	MBH-PHVA
⋮	⋮	⋮	⋮	⋮
D.[N _{free} low] ^{MBH}	MBH-PHVA			
		MBH	PHVA	MBH-PHVA

Figure 4. Schematic representation of the FHVA and PHVA cases considered in this work. Fixed parts of the system are indicated black, while the mobile blocks are shaded. Case A represents the FHVA method, while B.[N_{free} high] (almost all atoms free) to B.[N_{free} low] (almost all atoms fixed) corresponds to all PHVA schemes considered. Cases C and D are similar to A and B but with a mobile block description of the hydrocarbon (see text).

acid site to finally end up with a zeolite unit cell for which all zeolite atoms are considered free. Details on the various PHVA schemes used are given in Supporting Information.

The PHVA choices are labeled by the notation $[N_{\text{free}}^{\text{zeo}}, N_{\text{fixed}}^{\text{zeo}}]$ indicating the number of free ($N_{\text{free}}^{\text{zeo}}$) and fixed ($N_{\text{fixed}}^{\text{zeo}}$) zeolite atoms; all hydrocarbon atoms are free in case of FHVA or PHVA. An index (a, b, or c) is added if different schemes exist with the same number of free and fixed zeolite atoms but with different selections of atoms. For example, for H-ZSM-22 (see Section 3) in the [1,108] scheme, 1 zeolite atom (in this case the acid proton) is free, and the other 108 H-ZSM-22 zeolite atoms are fixed. Also for H-ZSM-22, in scheme [43,66], 43 zeolite atoms are free, while 66 atoms are fixed. Note that the schemes [109,0] and [145,0], respectively, H-ZSM-22 and H-FAU (see Section 3), correspond to the FHVA case (all atoms free, 0 atoms fixed). When MBH-FHVA or MBH-PHVA is applied, in which the hydrocarbon is modeled as a mobile rigid block, the scheme is denoted as $[N_{\text{free}}^{\text{zeo}}, N_{\text{fixed}}^{\text{zeo}}]^{\text{MBH}}$.

3. COMPUTATIONAL DETAILS

Investigated Systems. Physisorption of n-octane and of isobutene and chemisorption of isobutene in both H-ZSM-22

Table 1. Different PHVA Schemes Considered in This Study for H-ZSM-22 and H-FAU^a

H-ZSM-22 PHVA scheme	$N_{\text{free}}^{\text{zeo}}$ (T_{free})		$N_{\text{fixed}}^{\text{zeo}}$ (T_{fixed})	
[1,108]	1	(0)	108	(36)
[8,101]	8	(4)	101	(32)
[16,93]	16	(8)	93	(28)
[28,81]	28	(14)	81	(22)
[43,66] ^a	43	(6)	66	(30)
[43,66] ^b	43	(20)	66	(16)
[58,51]	58	(22)	51	(14)
[67,42]	67	(30)	42	(6)
[82,27]	82	(32)	27	(4)
[106,3]	106	(33)	3	(3)
[107,2] ^a	107	(33)	2	(3)
[107,2] ^b	107	(33)	2	(3)
[107,2] ^c	107	(34)	2	(2)
[108,1] ^a	108	(35)	1	(1)
[108,1] ^b	108	(35)	1	(1)
[108,1] ^c	108	(36)	1	(0)
[109,0]	109	(36)	0	(0)
H-FAU PHVA scheme	$N_{\text{free}}^{\text{zeo}}$ (T_{free})		$N_{\text{fixed}}^{\text{zeo}}$ (T_{fixed})	
[1,144]	1	(0)	144	(48)
[6,139]	6	(3)	139	(45)
[19,126]	19	(8)	126	(40)
[26,119]	26	(11)	119	(37)
[38,107]	38	(16)	107	(32)
[55,90]	55	(24)	90	(24)
[100,45]	100	(40)	45	(8)
[105,40]	105	(40)	40	(8)
[142,3]	142	(45)	3	(3)
[143,2] ^a	143	(46)	2	(2)
[143,2] ^b	143	(46)	2	(2)
[143,2] ^c	143	(48)	2	(0)
[144,1] ^a	144	(47)	1	(1)
[144,1] ^b	144	(47)	1	(1)
[144,1] ^c	144	(48)	1	(0)
[145,0]	145	(48)	0	(0)

^aThe different schemes are labeled as $[N_{\text{free}}^{\text{zeo}}, N_{\text{fixed}}^{\text{zeo}}]$, mentioning the number of free zeolite atoms ($N_{\text{free}}^{\text{zeo}}$) and the number of fixed ($N_{\text{fixed}}^{\text{zeo}}$) zeolite atoms; the number of free or fixed T (Al or Si) atoms is mentioned between brackets.

and H-FAU have been studied. The unit cells of H-ZSM-22 and H-FAU are sufficiently small to compute the benchmark full Hessian at a reasonable computational cost. H-ZSM-22 is a medium-pore zeolite characterized by 10-membered ring channels and has a cell composition $\text{Si}_{33}\text{AlO}_{72}\text{H}$. The optimized unit cell parameters are $a = 1132.9$ pm, $b = 1130.8$ pm, $c = 1542.8$ pm, $\alpha = 90.14^\circ$, $\beta = 90.07^\circ$, and $\gamma = 77.04^\circ$. H-FAU is a large-pore zeolite characterized by supercages connected by 12-membered rings, has a cell composition $\text{Si}_{47}\text{AlO}_{96}\text{H}$, and is characterized by the unit cell parameters $a = 1740.1$ pm, $b = 1736.2$ pm, $c = 1742.4$ pm, $\alpha = 59.88^\circ$, $\beta = 59.82^\circ$, and $\gamma = 59.87^\circ$. Calculated physisorption entropies for n-octane and isobutene in H-ZSM-22 and H-FAU are compared with experimental and simulation data from literature.^{9,10,44–46} Figure 1 depicts the different studied adsorption complexes in H-ZSM-22 and H-FAU.

Table 2. Geometry Optimization Criteria for the Zeolite, the Physisorbed *n*-Octane and Isobutene, and the Chemisorbed *t*-Butyl Carbenium Ion, *i*-Butoxy Alkoxide and *t*-Butoxy Alkoxide in H-ZSM-22 and H-FAU^a

	H-ZSM-22			H-FAU		
	E_{cutoff} (eV)	ΔE_{SCF} (eV)	max force (eV/Å)	E_{cutoff} (eV)	ΔE_{SCF} (eV)	max force (eV/Å)
zeolite	400	10^{-8}	0.010	400	10^{-8}	0.010
<i>n</i> -octane	600	10^{-10}	0.010	—	—	—
<i>i</i> -butene	600	10^{-10}	0.015	400	10^{-8}	0.015
<i>t</i> -butyl	400	10^{-8}	0.015	—	—	—
<i>i</i> -butoxy	400	10^{-8}	0.020	400	10^{-10}	0.010
<i>t</i> -butoxy	400	10^{-8}	0.020	400	10^{-8}	0.015

^a The used plane-wave energy cutoff E_{cutoff} , the SCF loop convergence criterion ΔE_{SCF} , and the maximum force on the atoms are given.

Periodic DFT Calculations. Periodic DFT calculations are performed using VASP.^{33–36} The total energy is calculated solving the Kohn–Sham equations of DFT using the gradient-corrected functionals of Perdew and Wang 91 (PW91) as the exchange–correlation functional.⁴⁷ The calculations are performed using the projector-augmented wave (PAW) method. This method was originally developed by Blöchl⁴⁸ and was adapted by Kresse and Joubert.⁴⁹ In general, a plane-wave cutoff of 400 eV was used in the calculations, and the Brillouin zone sampling was restricted to the Γ point.

Adsorption complexes are optimized in two steps. First, the conjugate gradient minimization algorithm was used to obtain a preoptimized structure until forces dropped below 0.05 eV/Å. Second, a quasi-Newton algorithm is employed to further optimize the structure until the maximum forces on the atoms are at least lower than 0.02 eV/Å. The loop for solving the electronic self-consistent field (SCF) equations iteratively is stopped when the difference between two consecutive energies is such that the accuracy of the energy is of the order of 10^{-8} eV.

After optimization of the adsorption complexes, a FHVA has been performed for all structures: The Hessian is calculated numerically by imposing positive and negative displacements in the x , y , and z -directions ($\Delta = 0.015$ Å) on each atom. We always carefully checked that no unwanted imaginary frequencies were present, ensuring a minimum-energy structure. However, especially in the case of loosely bonded complexes, such as physisorption of *n*-octane and isobutene, unwanted imaginary frequencies may appear when using the above-mentioned standard optimization settings (energy cutoff 400 eV, maximum force 0.02 eV/Å, and SCF loop convergence criteria 10^{-8} eV). In these cases, tighter VASP settings were necessary to get rid of the spurious imaginary frequencies: (i) maximum force on the atoms down to 0.01 eV/Å, (ii) SCF loop convergence criteria down to 10^{-10} eV, and (iii) plane-wave energy cutoff up to 600 eV. Especially the increase of the plane wave energy cutoff was found to be crucial to ensure the absence of unwanted imaginary frequencies. Table 2 summarizes the used settings for the optimization of each of the studied structures. In the statistical thermodynamics postprocessing, no frequency scaling factor was used for the calculation of the physisorption and chemisorption entropies.

For the loosely bonded *n*-octane and *t*-butyl carbenium ion complexes in H-FAU, spurious imaginary frequencies were still present even when using the most stringent optimization criteria. Many different configurations of the hydrocarbon inside the H-FAU zeolite have been considered, and even reoptimization of the structures using a cutoff energy of 600 eV, SCF loop convergence criteria of 10^{-12} eV, and maximum forces on the

atoms of 0.004 eV/Å was unsuccessful. The potential energy surface of such loosely bonded adsorbates is very flat, which makes the calculations extremely sensitive to numerical noise. More details on these calculations can be found in Supporting Information. In these cases of unwanted imaginary frequencies, no further processing of the Hessian was performed, and therefore, the physisorbed *n*-octane and the chemisorbed *t*-butyl carbenium ion complexes in H-FAU are not further investigated in this study.

4. RESULTS AND DISCUSSION

As mentioned in the Introduction, the focus of this work is on physisorption and chemisorption entropies and not on physisorption and chemisorption enthalpies. First, FHVA and PHVA results are compared assuming an immobile adsorbate, i.e., the approach in which all harmonic frequencies also those corresponding to translation and rotation, are retained. Next, the mobile adsorbate method and the application of the MBH method for an unambiguous determination of translational and rotational frequencies are discussed for the loosely bonded complexes. Finally, inspired by the conclusions of this work, some general guidelines for performing vibrational analysis of extended systems are presented.

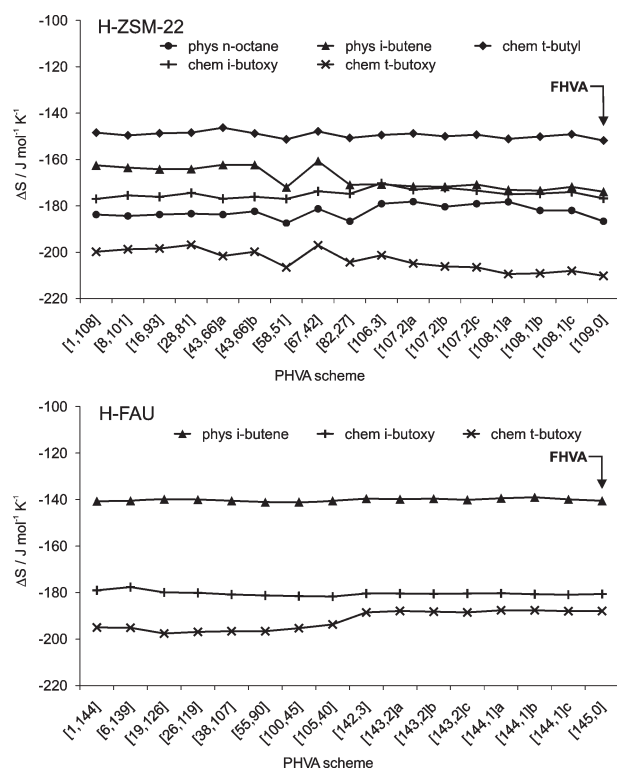
For completeness, the calculated physisorption and chemisorption enthalpies at 300 K for the studied complexes can be found in Supporting Information (Table S.1). Note that these enthalpy differences are independent of the PHVA scheme applied and equal to the FHVA results.

FHVA versus PHVA Assuming an Immobile Adsorbate. *FHVA Zero Frequencies.* In principle, FHVA calculations should yield three zero frequencies corresponding to the translation of the zeolite unit cell as a whole. In practice, very low frequencies are found for these motions of the unit cells, and the obtained values can be regarded as indication for the quality of the Hessian. Table 3 shows that the obtained imaginary frequencies have values between -1.5 and -6.0 cm^{-1} indicative of a well-conditioned Hessian considering the numerical calculation as implemented in VASP.

FHVA and PHVA Adsorption Entropies. Figure 5 shows the calculated physisorption and chemisorption entropies in H-ZSM-22 and H-FAU, respectively, assuming an immobile adsorbate for FHVA and the various PHVA schemes. The numerical values are given in the Supporting Information (Tables S.2 and S.3). In Figure 5, the PHVA scheme with only one free zeolite atom, i.e., the zeolite acidic H-atom, scheme [1,108] in H-ZSM-22 and scheme [1,144] in H-FAU, is found at the most left and the FHVA benchmark result at the most right of the abscis.

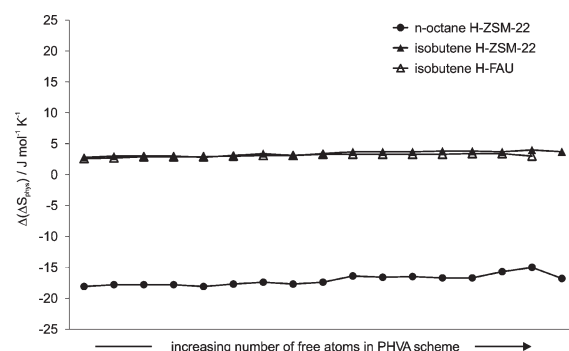
Table 3. Zero Frequencies Corresponding to the Translation of the Zeolite Unit Cell as an Indication of the Quality of the Hessian^a

zero frequencies (cm ⁻¹)	H-ZSM-22			H-FAU		
zeolite	-2.4	-1.9	-1.6	-2.3	-2.1	-1.8
phys n-octane	-4.3	-2.0	-1.5	—	—	—
phys <i>i</i> -butene	-5.5	-4.5	-3.4	-2.8	-1.3	-1.3
chem <i>i</i> -butoxy	-6.4	-5.3	-3.1	-1.4	-1.4	-1.3
chem <i>t</i> -butoxy	-5.1	-3.8	-3.5	-3.6	-2.5	-2.0
chem <i>t</i> -butyl	-6.0	-4.6	-3.6	—	—	—

^aImaginary frequencies are denoted with a minus sign.**Figure 5.** Physisorption and chemisorption entropies of n-octane, isobutene, *t*-butyl carbenium ion, *i*-butoxy and/or *t*-butoxy in H-ZSM-22 (top) and H-FAU (bottom) calculated with the immobile adsorbate method for FHVA and various PHVA schemes.

FHVA physisorption entropies of, respectively, n-octane and isobutene in H-ZSM-22 amount to -189 and -174 J mol⁻¹ K⁻¹, while FHVA chemisorption entropies for the *t*-butyl carbenium ion, and the *i*-butoxy, and *t*-butoxy alkoxides are -152 , -177 , and -210 J mol⁻¹ K⁻¹. In H-FAU on the other hand, FHVA physisorption and chemisorption entropies of isobutene, *i*-butoxy, and *t*-butoxy, respectively, are -141 , -181 , and -188 J mol⁻¹ K⁻¹. Entropy losses in H-FAU are somewhat lower as compared to H-ZSM-22 in accordance with the more open structure of H-FAU.

Comparison of FHVA with physisorption and chemisorption entropies calculated with the various PHVA schemes shows that all FHVA and PHVA results fall within a narrow range of approximately 10 J mol⁻¹ K⁻¹ for H-ZSM-22 as well as for H-FAU. In addition, no systematic trends in the deviation are

**Figure 6.** Deviations $\Delta(\Delta S_{\text{phys}})$ caused by neglecting the internal vibrations of the hydrocarbon for the FHVA and various PHVA schemes.

observed with increasing PHVA scheme size. The tighter convergence criteria for loosely bonded physisorption complexes (see Table 2) are needed to ensure this relatively small range. For instance, optimization of the n-octane and isobutene physisorption complexes in H-ZSM-22 using the standard settings (energy cutoff 400 eV, SCF convergence criterion 10^{-8} eV) lead to differences of physisorption entropies between FHVA and PHVA up to 30 J mol⁻¹ K⁻¹ (see Figure S.1, Supporting Information). This can be understood from the presence of floppy modes for the loosely bonded complexes: The small energy changes upon atom displacements result in low-lying frequencies which are easily contaminated with noise if less stringent convergence criteria are used for the plane-wave energy cutoff and the SCF convergence criterion.

The most striking conclusion however is that a PHVA calculation in its most simplified form, i.e., with only the zeolite acid H-atom considered free, yields similar physisorption and chemisorption entropies than a computationally much more expensive FHVA calculation. This implies that the internal zeolite vibrations have only a limited impact to the adsorption entropy, whereas the main contribution stems from the frequencies describing the internal vibrations in the hydrocarbon complex and/or the vibrations of the hydrocarbon relative to the zeolite. Apparently, only a small part of the extended zeolite catalyst needs to be considered for a high-quality vibrational analysis, yielding results comparable to a FHVA. This observation opens perspectives for other extended systems beyond zeolites, provided the simulation software allows the calculation of partial Hessians. The geometry of large systems can be fully or partially optimized at a high level of theory, while the vibrational analysis can be limited to subset of atoms, making this computationally demanding step considerably more feasible. The speedup can be roughly estimated to be

$$N/N_{\text{free}} \quad (2)$$

in which N and N_{free} respectively, are the total number of atoms and the number of free atoms in the unit cell, obviously accounting for the zeolite as well as the hydrocarbon atoms.

Influence of the Internal Vibrations of the Hydrocarbon. The influence of the internal hydrocarbon vibrations on the physisorption entropies of n-octane and isobutene in H-ZSM-22 and H-FAU is evaluated by comparing cases A with C and cases B.[i] with D.[i] (see Figure 4). In cases C and D, the hydrocarbon is considered as a mobile block such that its internal vibrations are absent in these models. Figure 6 shows the calculated differences

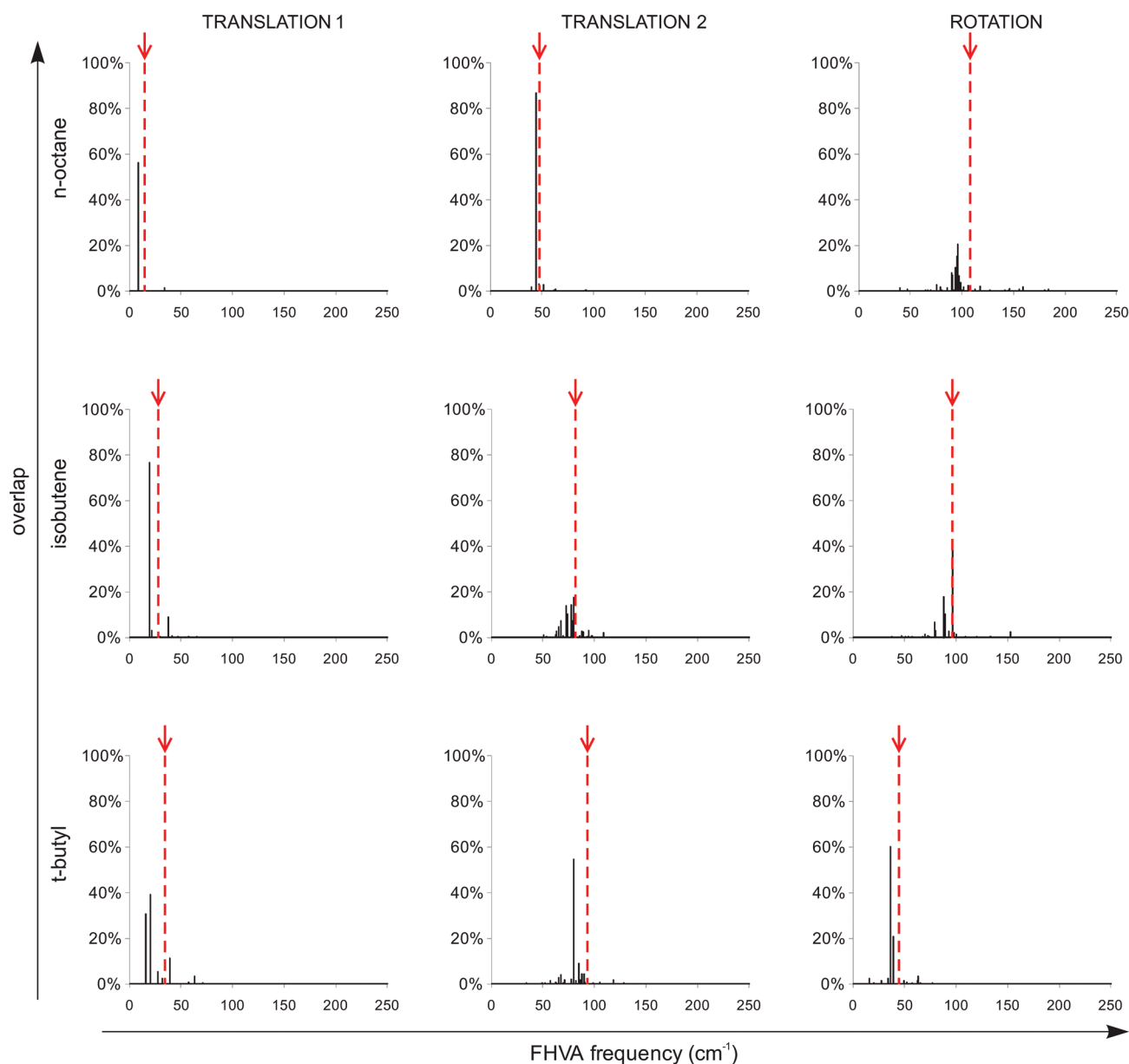


Figure 7. Overlap for the two translational modes and one rotational mode for *n*-octane (top), isobutene (middle), and *t*-butyl carbenium (bottom) in H-ZSM-22, as obtained from $[1,108]^{MBH}$, with the FHVA calculated modes (scheme $[109,0]$) in H-ZSM-22. The red dashed line, indicated by the arrow, represents the unique translational, respectively, rotational frequency obtained from the $[1,108]^{MBH}$ calculation. The overlap of the corresponding mode with the FHVA modes is indicated by the black lines spectrum.

$\Delta(\Delta S_{\text{phys}})$ of the physisorption entropies, caused by neglecting the internal hydrocarbon vibration. For FHVA, $\Delta(\Delta S_{\text{phys}})$ is calculated as follows:

$$\Delta(\Delta S_{\text{phys}}) = \Delta S_{\text{phys}}^{\text{case A}} - \Delta S_{\text{phys}}^{\text{case C}} \quad (3)$$

In case of PHVA, $\Delta(\Delta S_{\text{phys}})$ is calculated as

$$\Delta(\Delta S_{\text{phys}}) = \Delta S_{\text{phys}}^{\text{case B.[i]}} - \Delta S_{\text{phys}}^{\text{case D.[i]}} \quad (4)$$

Figure 6 shows that the internal vibrational entropy of isobutene has only a minor contribution to the adsorption entropy since the calculated differences $\Delta(\Delta S_{\text{phys}})$ amount to about 3–4 J mol⁻¹ K⁻¹ in H-ZSM-22 as well as in H-FAU. Gas-

phase isobutene is a rather small molecule, and no frequencies lower than 100 cm⁻¹, which contribute most to the entropy, are present. The lowest frequencies of 173 and 208 cm⁻¹ correspond to the (anti)symmetric methyl rotation. Consequently, considering isobutene as a mobile block has only a minor influence on the physisorption entropy. In contrast, gas-phase *n*-octane is much more flexible and subject to many low-frequency motions; 3 frequencies are lower than 100 cm⁻¹, the most pronounced one being an internal rotation at 43 cm⁻¹. Since these internal rotations are rather hindered when the *n*-octane molecule is physisorbed in the zeolite pores, the contributions to the entropy do not cancel out, and $\Delta(\Delta S_{\text{phys}})$ amounts to -14 to -18 J mol⁻¹ K⁻¹ in the case of *n*-octane adsorption in H-ZSM-22. It can thus be expected that for hydrocarbons with many possible

Table 4. Immobile versus Mobile Physisorption and Chemisorption Entropies of n-Octane, Isobutene, and *t*-Butyl Carbenium Ion Calculated with FHVA^a

(J mol ⁻¹ K ⁻¹)	H-ZSM-22		H-FAU		gas phase	
	immobile	mobile	immobile	mobile	3D trans	3D rot
n-octane	-189	-148	–	–	168	117
isobutene	-174	-126	-141	-94	159	96
<i>t</i> -butyl	-152	-102	–	–	159	104

^a The translational and rotational contributions to the entropy of the gas-phase hydrocarbons are added as a reference.

internal rotations in gas phase, such as long n-alkanes (like n-octane in this case), MBH-FHVA (C) or MBH-PHVA (D.[i]) calculations yield significantly different physisorption entropies as compared to FHVA (A) or PHVA (B.[i]).

Immobile versus Mobile Adsorbate. *MBH-PHVA-Based Selection of Translational and Rotational Frequencies – overlap with FHVA spectrum.* As explained in Section 2, MBH-PHVA schemes [1,108]^{MBH} in H-ZSM-22 and [1,144]^{MBH} in H-FAU are applied for the identification of translational and rotational frequencies of loosely bonded complexes. The overlap of an MBH-PHVA mode with an FHVA mode is defined as the absolute value of the dot product of the normalized MBH-PHVA mode with the normalized FHVA mode, such that it lies in the range 0–1. A high overlap value indicates that the MBH-PHVA and the FHVA modes represent similar motions. Since the MBH-PHVA modes represent translational and rotational motions of the hydrocarbon in the zeolite pore, overlaps of the MBH-PHVA modes with all FHVA modes are adequate parameters to identify the translational and rotational motions in the FHVA spectrum. This is shown in Figure 7 for physisorption of n-octane and isobutene and chemisorption of the *t*-butyl carbenium ion in H-ZSM-22. The dashed line, indicated by the arrow, represents the unique translational or rotational frequency obtained from the MBH-PHVA calculation. The black line spectrum illustrates the mixing of the translational and rotational modes over several frequencies in the FHVA calculation, also revealing the ambiguity when translational and rotational frequencies are selected manually from the FHVA calculations. A similar figure for the physisorption of isobutene in H-FAU can be found in the Supporting Information (Figure S.2). Figure 7 also shows that the unique translational and rotational frequencies of the MBH-PHVA approach correspond reasonably well to the frequency spectrum obtained from the FHVA calculations. The frequency values are slightly higher, as expected in view of the large part of the zeolite that is kept fixed in the MBH-PHVA calculation, thus constraining the remaining motions.

FHVA and PHVA Adsorption Entropies. Table 4 compares the immobile and mobile adsorbate method for the FHVA calculation of the physisorption entropies of n-octane and isobutene and the chemisorption entropy of the *t*-butyl carbenium ion. The 3D and 3D rotational entropies of the molecules in the gas phase are added, since the gas-phase translational/rotational entropy can be regarded as a rough (but incorrect) estimate for the adsorption entropies.⁹

The mobile adsorbate method as described in eq 1 has been applied, using the translational and rotational frequencies obtained from the MBH-PHVA approach explained in the previous paragraph. Because of the replacement of vibrational

Table 5. Comparison of Our Calculated Values for the Physisorption Entropy of n-Octane (Mobile Adsorbate) with Values Reported in Literature

(J mol ⁻¹ K ⁻¹)	H-ZSM-22	H-FAU
n-octane		
this work	-148	–
Denayer et al. ^{45,46a}	-159	-83 to -98
Eder et al. ⁴⁴	–	-101
De Moor et al. ⁹	–	-88
isobutene		
this work	-126	-94
De Moor et al. ⁹	–	-100

^a Values have been revisited, see De Moor et al.¹⁰

contributions by free translational and free rotational contributions, entropy losses upon adsorption become smaller. A shift of about 50 J mol⁻¹ K⁻¹ is observed for the entropies, assuming an immobile versus a mobile adsorbate. Mobile adsorbate physisorption and chemisorption entropies for n-octane, isobutene, and *t*-butyl carbenium ion amount to -148, -126, and -102 J mol⁻¹ K⁻¹ in H-ZSM-22, while the physisorption entropy for isobutene amounts to -94 J mol⁻¹ K⁻¹ in H-FAU. As already mentioned above, the entropy loss upon isobutene physisorption is somewhat higher in H-ZSM-22 than in H-FAU, due to the more confined structure of the former.

The immobile–mobile adsorbate entropy shift is independent of the FHVA or PHVA scheme, and as a consequence, PHVA and FHVA results are nearly identical and differ at most 10 J mol⁻¹ K⁻¹, as was also the case for the immobile adsorbate method. Fluctuations observed in the immobile adsorbate method are also present in the mobile adsorbate method (see Figure 5). There are no systematic trends, and PHVA schemes with only one free zeolite atom, [1,108] (H-ZSM-22) and [1,144] (H-FAU), yield values that are very similar to the FHVA results. An uncertainty of 10 J mol⁻¹ K⁻¹ on the physisorption or chemisorption entropy leads to an uncertainty of a factor 3 on the adsorption equilibrium coefficient, which is very acceptable. Figures presenting the effect of PHVA assuming a mobile adsorbate, analogous to Figure 5, are given in the Supporting Information (Figures S.3 and S.4) as well as the numerical values for the physisorption and chemisorption entropies (Tables S.2 and S.3, Supporting Information).

Comparison with Experiment. In Table 5, FHVA physisorption entropies for n-octane and isobutene in H-ZSM-22 and H-FAU, assuming a mobile adsorbate, are compared with experimental values and values reported for other molecular simulations from literature. A very good agreement is observed between the VASP physisorption entropies calculated in this work and the experimental values determined by Denayer et al.^{45,46} Our VASP calculated physisorption entropies also nicely agree with QM-Pot(MP2//B3LYP) physisorption entropies reported by De Moor et al.⁹ The good agreement between these two different simulation methods for the physisorption of isobutene in H-FAU indicates that the influence of the van der Waals stabilizing interactions—these are adequately described by QM-Pot(MP2//B3LYP) but not by VASP—on the value of the entropies is rather limited for the zeolite systems under study.

Differences between our calculated physisorption entropies and the literature data do not exceed 10–15 J mol⁻¹ K⁻¹. One of the possible explanations for the differences between the simulation and the experimental results may relate to the improper

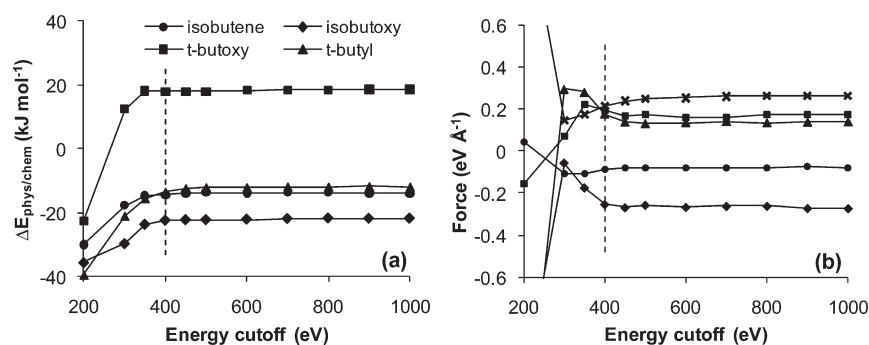


Figure 8. Influence of energy cutoff (a) on the physisorption/chemisorption energy and (b) on the gradient of some arbitrarily chosen atoms of a nonoptimized isobutene, *i*-butoxy, *t*-butoxy, and *t*-butyl complexes in H-ZSM-22.

description of internal frustrated rotational modes by the harmonic oscillator approximation; such a refinement of the approach can be the topic of a further thorough study.

Guidelines for Vibrational Analysis of Extended Systems.

From our results, we can state that the use of PHVA (indicated by the dashed line in Figure 2) for zeolite systems, and by extension for other catalytic systems, leads to a significant and structural reduction of the computational resources needed. Evidently, this advantageous use of PHVA is only possible with simulation packages in which the partial Hessian calculation is implemented, as is the case in VASP.^{33–36} For loosely bonded complexes, such as physisorbed alkanes or alkenes and carbenium ions, the mobile adsorbate method is proposed, with an unambiguous selection of the translational or rotational modes based on an MBH–PHVA calculation using TAMkin.^{37,38} As discussed below, a well-conditioned Hessian is however in any case required.

Based on this study on the physisorption and chemisorptions of *n*-octane and isobutene in H-ZSM-22 and H-FAU, some guidelines can be formulated for performing geometry optimization and vibrational analysis on extended zeolite systems using VASP.^{33–36} Strategies and settings for electronic energy calculations of extended zeolite systems are well established. However, commonly used convergence criteria sufficiently accurate for a geometry optimization appear not to be sufficiently tight for successful normal-mode analysis; frequently, unwanted imaginary frequencies are present, in particular for loosely bonded physisorption complexes. Stricter optimization criteria are thus necessary to construct a well-conditioned Hessian, further compounding the computational cost of the numerical Hessian calculations. As illustrated in this study, this drawback can be largely overcome by using PHVA schemes. Indeed, it has been shown that for the systems considered in this study, even if only the zeolite acid H-atom is considered to be free, the PHVA results are very similar to the computationally much more expensive FHVA results. A deviation from the FHVA adsorption entropy of $10 \text{ J mol}^{-1} \text{ K}^{-1}$ at most has been found. This is a very important conclusion, in view of the fact that a full vibrational analysis is in general much more time-consuming than a geometry optimization using more stringent convergence criteria. From the theoretical speed-up that can be calculated using eq 2, it is concluded that the [1,108] PHVA scheme is some 9 times faster than FHVA for isobutene physisorption in H-ZSM-22, i.e., the computational cost is reduced by 89%. In H-FAU, the [1,144] PHVA scheme is even 12 times faster, leading to a reduction of the computational cost of 92%. Illustrating figures can be found in the Supporting Information. Clearly, these schemes are computationally very

attractive for normal-mode analysis of extended zeolite structures.

Although for studies of zeolite systems using VASP an energy cutoff of 400 eV is widely accepted, we have shown that a cutoff of 600 eV was necessary for some loosely bonded physisorption complexes. Figure 8 shows that convergence of the energy differences and gradients starts at an energy cutoff of 400 eV and explains the need for an energy cutoff of 600 eV in some of the studied systems. Figure S.6 in Supporting Information shows the convergence of the gradient as function of the plane-wave energy cutoff for a larger number of atoms. In general, it can be stated that obtaining a well-conditioned Hessian requires that the value of the plane-wave energy cutoff is set large enough to ensure that energy differences and forces have converged and that an energy cutoff of 400 eV is on the edge of what is required to calculate reliable adsorption entropies in extended zeolite systems.

Also, other extended catalytic systems than zeolites may benefit from this study. As a general guideline, we propose to check the convergence of the energy differences and forces as function of the plane-wave energy cutoff (via some single point calculations) in order to estimate the minimum energy cutoff needed for a particular system, before effectively running the Hessian calculations in VASP. Another option is to follow a similar methodology as presented here for evaluating and comparing FHVA and (various) PHVA results using TAMkin.^{37,38}

5. CONCLUSIONS

Physisorption and chemisorption of *n*-octane and isobutene complexes in H-ZSM-22 and H-FAU have been studied using periodic DFT calculations. A computationally efficient procedure for performing normal-mode analysis in extended zeolite systems is presented. Physisorption and chemisorption entropies have been calculated from FHVA and various PHVA schemes. All PHVA evaluations can be performed using the TAMkin program, based on the FHVA Hessian without any additional computational cost. The agreement between FHVA and PHVA results is satisfactory, as differences in the calculated physisorption and chemisorption entropies do not exceed $10 \text{ J mol}^{-1} \text{ K}^{-1}$ provided that stricter convergence criteria for optimization are used, especially for loosely bonded complexes in zeolites. Hence, PHVA provides an attractive alternative for the computationally demanding FHVA calculations to obtain reasonably accurate entropies. The reduction of computational cost when performing

a PHVA instead of a FHVA is significant and amounts to 1 order of magnitude for the systems in this study.

An unambiguous method is presented for the identification of hydrocarbon rotational and translation modes relative to the zeolite. These modes are needed for the application of the mobile adsorbate method in case of loosely bonded complexes. The vibrational frequencies corresponding to these translational and rotational modes that are replaced by free translational and rotational contributions are easily identified based on a MBH–PHVA calculation using the TAMkin package. Physisorption entropies of n-octane and isobutene in H-ZSM-22 and H-FAU obtained from the mobile adsorbate method are predicted within $10\text{--}15\text{ J mol}^{-1}\text{ K}^{-1}$ as compared to experimental and simulated data available in literature.

The (MBH–)PHVA procedure presented in this work is directly applicable to calculate activation entropies and reaction entropies in extended zeolite systems. The procedure also provides a more efficient methodology for normal vibrational analysis in other extended (catalytic) systems.

■ ASSOCIATED CONTENT

S Supporting Information. Discussion of the loosely bonded n-octane and *t*-butyl carbenium ion complexes in H-FAU, for which unwanted imaginary frequencies are found. Values for physisorption and chemisorption enthalpies and entropies are mentioned, and details on the PHVA choices in H-ZSM-22 and H-FAU are given. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: MarieFrancoise.Reyniers@ugent.be.

■ ACKNOWLEDGMENT

This work is supported by the Long Term Structural Methusalem Funding by the Flemish Government—grant number BOF09/01M00409, the FWO (Fund for Scientific Research Flanders), the BELSPO (Belgian Federal Science Policy Office in the frame of IAP/6/27), the E.C. (Network of Excellence IDECAT, NMP3-CT-2005-011730) and by the BOF (Research Fund of Ghent University). The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by Ghent University. V.V.S. acknowledges the European Research Council under the European Community's Seventh Framework Programme (FP7(2007-2013) ERC grant agreement number 240483).

■ REFERENCES

- (1) Smit, B.; Maesen, T. L. M. Molecular Simulations of Zeolites: Adsorption, Diffusion, and Shape Selectivity. *Chem. Rev.* **2008**, *108*, 4125–4184.
- (2) Pascual, P.; Ungerer, P.; Tavitian, B.; Pernot, P.; Boutin, A. Development of a transferable guest-host force field for adsorption of hydrocarbons in zeolites - I. Reinvestigation of alkane adsorption in silicalite by grand canonical Monte Carlo simulation. *Phys. Chem. Chem. Phys.* **2003**, *5*, 3684–3693.
- (3) Clark, L. A.; Sierka, M.; Sauer, J. Stable mechanistically-relevant aromatic-based carbenium ions in zeolite catalysts. *J. Am. Chem. Soc.* **2003**, *125*, 2136–2141.

- (4) Clark, L. A.; Sierka, M.; Sauer, J. Computational elucidation of the transition state shape selectivity phenomenon. *J. Am. Chem. Soc.* **2004**, *126*, 936–947.
- (5) Nieminen, V.; Sierka, M.; Murzin, D. Y.; Sauer, J. Stabilities of C-3-C-5 alkoxide species inside H-FER zeolite: a hybrid QM/MM study. *J. Catal.* **2005**, *231*, 393–404.
- (6) Pantu, P.; Boekfa, B.; Limtrakul, J. The adsorption of saturated and unsaturated hydrocarbons on nanostructured zeolites (H-MOR and H-FAU): An ONIOM study. *J. Mol. Catal. A: Chem.* **2007**, *277*, 171–179.
- (7) Boronat, M.; Corma, A. Are carbenium and carbonium ions reaction intermediates in zeolite-catalyzed reactions?. *Appl. Catal., A* **2008**, *336*, 2–10.
- (8) De Moor, B. A.; Reyniers, M. F.; Sierka, M.; Sauer, J.; Marin, G. B. Physisorption and Chemisorption of Hydrocarbons in H-FAU Using QM-Pot(MP2//B3LYP) Calculations. *J. Phys. Chem. C* **2008**, *112*, 11796–11812.
- (9) De Moor, B. A.; Reyniers, M. F.; Marin, G. B. Physisorption and chemisorption of alkanes and alkenes in H-FAU: a combined ab initio-statistical thermodynamics study. *Phys. Chem. Chem. Phys.* **2009**, *11*, 2939–2958.
- (10) De Moor, B. A.; Reyniers, M. F.; Gobin, O. C.; Lercher, J. A.; Marin, G. B. Adsorption of C2-C8 n-alkanes in zeolites. *J. Phys. Chem. C* **2011**, *115*, 1204–1219.
- (11) Tuma, C.; Sauer, J. Treating dispersion effects in extended systems by hybrid MP2: DFT calculations - protonation of isobutene in zeolite ferrierite. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3955–3965.
- (12) Svelle, S.; Tuma, C.; Rozanska, X.; Kerber, T.; Sauer, J. Quantum Chemical Modeling of Zeolite-Catalyzed Methylation Reactions: Toward Chemical Accuracy for Barriers. *J. Am. Chem. Soc.* **2009**, *131*, 816–825.
- (13) Mccann, D. M.; Lesthaeghe, D.; Kletnieks, P. W.; Guenther, D. R.; Hayman, M. J.; Van Speybroeck, V.; Waroquier, M.; Haw, J. F. A complete catalytic cycle for supramolecular methanol-to-olefins conversion by linking theory with experiment. *Angew. Chem., Int. Ed.* **2008**, *47*, 5179–5182.
- (14) Lesthaeghe, D.; Horre, A.; Waroquier, M.; Marin, G. B.; Van Speybroeck, V. Theoretical Insights on Methylbenzene Side-Chain Growth in ZSM-5 Zeolites for Methanol-to-Olefin Conversion. *Chem.—Eur. J.* **2009**, *15*, 10803–10808.
- (15) Vandichel, M.; Lesthaeghe, D.; Van der Mynsbrugge, J.; Waroquier, M.; Van Speybroeck, V. Assembly of cyclic hydrocarbons from ethene and propene in acid zeolite catalysis to produce active catalytic sites for MTO conversion. *J. Catal.* **2010**, *271*, 67–78.
- (16) Rozanska, X.; Demuth, T.; Hutschka, F.; Hafner, J.; van Santen, R. A. A periodic structure density functional theory study of propylene chemisorption in acidic chabazite: Effect of zeolite structure relaxation. *J. Phys. Chem. B* **2002**, *106*, 3248–3254.
- (17) Rozanska, X.; van Santen, R. A.; Demuth, T.; Hutschka, F.; Hafner, J. A periodic DFT study of isobutene chemisorption in proton-exchanged zeolites: Dependence of reactivity on the zeolite framework structure. *J. Phys. Chem. B* **2003**, *107*, 1309–1315.
- (18) Benco, L.; Hafner, J.; Hutschka, F.; Toulhoat, H. Physisorption and chemisorption of some n-hydrocarbons at the Bronsted acid site in zeolites 12-membered ring main channels: Ab initio study of the gmelinite structure. *J. Phys. Chem. B* **2003**, *107*, 9756–9762.
- (19) Demuth, T.; Rozanska, X.; Benco, L.; Hafner, J.; van Santen, R. A.; Toulhoat, H. Catalytic isomerization of 2-pentene in H-ZSM-22 - A DFT investigation. *J. Catal.* **2003**, *214*, 68–77.
- (20) Tuma, C.; Sauer, J. Protonated isobutene in zeolites: tert-butyl cation or alkoxide?. *Angew. Chem., Int. Ed.* **2005**, *44*, 4769–4771.
- (21) Kerber, T.; Sierka, M.; Sauer, J. Application of semiempirical long-range dispersion corrections to periodic systems in density functional theory. *J. Comput. Chem.* **2008**, *29*, 2088–2097.
- (22) Yaluri, G.; Rekoske, J. E.; Aparicio, L. M.; Madon, R. J.; Dumesic, J. A. Isobutane Cracking Over Y-Zeolites 0.1. Development of A Kinetic-Model. *J. Catal.* **1995**, *153*, 54–64.
- (23) Narasimhan, C. S. L.; Thybaut, J. W.; Marin, G. B.; Jacobs, P. A.; Martens, J. A.; Denayer, J. F.; Baron, G. V. Kinetic modeling of pore

mouth catalysis in the hydroconversion of n-octane on Pt-H-ZSM-22. *J. Catal.* **2003**, *220*, 399–413.

(24) Thybaut, J. W.; Narasimhan, C. S. L.; Marin, G. B.; Denayer, J. F. M.; Baron, G. V.; Jacobs, P. A.; Martens, J. A. Alkylcarbenium ion concentrations in zeolite pores during octane hydrocracking on Pt/H-USY zeolite. *Catal. Lett.* **2004**, *94*, 81–88.

(25) Calvin, M. D.; Head, J. D.; Jin, S. Q. Theoretically modelling the water bilayer on the Al(111) surface using cluster calculations. *Surf. Sci.* **1996**, *345*, 161–172.

(26) Head, J. D. Computation of vibrational frequencies for adsorbates on surfaces. *Int. J. Quantum Chem.* **1997**, *65*, 827–838.

(27) Head, J. D.; Shi, Y. Characterization of Fermi resonances in adsorbate vibrational spectra using cluster calculations: Methoxy adsorption on Al(111) and Cu(111). *Int. J. Quantum Chem.* **1999**, *75*, 815–820.

(28) Head, J. D. A vibrational analysis with fermi resonances for methoxy adsorption on Cu(111) using ab initio cluster calculations. *Int. J. Quantum Chem.* **2000**, *77*, 350–357.

(29) Jin, S. Q.; Head, J. D. Theoretical Investigation of Molecular Water-Adsorption on the Al(111) Surface. *Surf. Sci.* **1994**, *318*, 204–216.

(30) Li, H.; Jensen, J. H. Partial Hessian vibrational analysis: the localization of the molecular vibrational energy and entropy. *Theor. Chem. Acc.* **2002**, *107*, 211–219.

(31) Ghysels, A.; Van Neck, D.; Van Speybroeck, V.; Verstraelen, T.; Waroquier, M. Vibrational modes in partially optimized molecular systems. *J. Chem. Phys.* **2007**, *126*, Art.No.224102.

(32) Cramer, C. J. *Essentials of Computational Chemistry: Theories and Models*; 2nd ed.; John Wiley & Sons Ltd.: Chichester, England, 2005.

(33) Kresse, G.; Hafner, J. Abinitio Molecular-Dynamics for Liquid-Metals. *Phys. Rev. B* **1993**, *47*, 558–561.

(34) Kresse, G.; Hafner, J. Ab-Initio Molecular-Dynamics Simulation of the Liquid-Metal Amorphous-Semiconductor Transition in Germanium. *Phys. Rev. B* **1994**, *49*, 14251–14269.

(35) Kresse, G.; Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **1996**, *54*, 11169–11186.

(36) Kresse, G.; Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **1996**, *6*, 15–50.

(37) Center of Molecular Modeling (CMM) of Ghent University; Ghent University: Ghent, Belgium; <http://molmod.ugent.be/code/wiki>. Accessed May 25, 2010.

(38) Ghysels, A.; Verstraelen, T.; Hemelsoet, K.; Van Speybroeck, V.; Waroquier, M. TAMkin: a versatile package for vibrational analysis and kinetics. *J. Chem. Inf. Model* **2010**, *50*, 1736–1750.

(39) Ghysels, A.; Van Neck, D.; Waroquier, M. Cartesian formulation of the mobile block Hessian approach to vibrational analysis in partially optimized systems. *J. Chem. Phys.* **2007**, *127*, no. 164108.

(40) Ghysels, A.; Van Speybroeck, V.; Verstraelen, T.; Van Neck, D.; Waroquier, M. Calculating reaction rates with partial Hessians: Validation of the mobile block Hessian approach. *J. Chem. Theory Comput.* **2008**, *4*, 614–625.

(41) Ghysels, A.; Van Speybroeck, V.; Pauwels, E.; Van Neck, D.; Brooks, B. R.; Waroquier, M. Mobile Block Hessian Approach with Adjoined Blocks: An Efficient Approach for the Calculation of Frequencies in Macromolecules. *J. Chem. Theory Comput* **2009**, *5*, 1203–1215.

(42) Ghysels, A.; Van Neck, D.; Brooks, B. R.; Van Speybroeck, V.; Waroquier, M. Normal modes for large molecules with arbitrary link constraints in the mobile block Hessian approach. *J. Chem. Phys.* **2009**, *130*, no. 084107.

(43) Ghysels, A.; Van Speybroeck, V.; Pauwels, E.; Catak, S.; Brooks, B. R.; Van Neck, D.; Waroquier, M. Comparative Study of Various Normal Mode Analysis Techniques Based on Partial Hessians. *J. Comput. Chem.* **2010**, *31*, 994–1007.

(44) Eder, F.; Stockenhuber, M.; Lercher, J. A. Bronsted acid site and pore controlled siting of alkane sorption in acidic molecular sieves. *J. Phys. Chem. B* **1997**, *101*, 5414–5419.

(45) Denayer, J. F.; Baron, G. V.; Martens, J. A.; Jacobs, P. A. Chromatographic study of adsorption of n-alkanes on zeolites at high temperatures. *J. Phys. Chem. B* **1998**, *102*, 3077–3081.

(46) Denayer, J. F.; Souverijns, W.; Jacobs, P. A.; Martens, J. A.; Baron, G. V. High-temperature low-pressure adsorption of branched C-5-C-8 alkanes on zeolite beta, ZSM-5, ZSM-22, zeolite Y, and mordenite. *J. Phys. Chem. B* **1998**, *102*, 4588–4597.

(47) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. Atoms, Molecules, Solids, and Surfaces - Applications of the Generalized Gradient Approximation for Exchange and Correlation. *Phys. Rev. B* **1992**, *46*, 6671–6687.

(48) Blöchl, P. E. Projector Augmented-Wave Method. *Phys. Rev. B* **1994**, *50*, 17953–17979.

(49) Kresse, G.; Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **1999**, *59*, 1758–1775.

First Principles-Based Calculations of Free Energy of Binding: Application to Ligand Binding in a Self-Assembling Superstructure

Stephen Fox,[†] Hannes G. Wallnoefer,[‡] Thomas Fox,[‡] Christofer S. Tautermann,[‡] and Chris-Kriton Skylaris^{*,†}

[†]School of Chemistry, University of Southampton, Southampton, Hampshire SO17 1BJ, United Kingdom

[‡]Department for Lead Identification and Optimization Support, Boehringer Ingelheim Pharma GmbH & Co KG, 88397 Biberach, Germany

ABSTRACT: The accurate prediction of ligand binding affinities to a protein remains a desirable goal of computational biochemistry. Many available methods use molecular mechanics (MM) to describe the system, however, MM force fields cannot fully describe the complex interactions involved in binding, specifically electron transfer and polarization. First principles approaches can fully account for these interactions, and with the development of linear-scaling first principles programs, it is now viable to apply first principles calculations to systems containing tens of thousands of atoms. In this paper, a quantum mechanical Poisson–Boltzmann surface area approach is applied to a model of a protein–ligand binding cavity, the “tennis ball” dimer. Results obtained from this approach demonstrate considerable improvement over conventional molecular mechanics Poisson–Boltzmann surface area due to the more accurate description of the interactions in the system. For the first principles calculations in this study, the linear-scaling density functional theory program ONETEP is used, allowing the approach to be applied to receptor–ligand complexes of pharmaceutical interest that typically include thousands of atoms.

1. INTRODUCTION

With the growing number of experimentally determined 3D molecular structures refined to a high atomic resolution, molecular modeling is expanding its role in understanding structure/function relationships of biomolecules. Techniques of increasing sophistication are available for describing atomic forces, ranging from classical molecular mechanics (MM) with coarse grained or atomistic force fields to first principles electronic structure calculations.¹ Computational simulations with these techniques can be used to calculate structural, dynamic, and thermodynamic properties and have found wide usage as tools for assessing potential pharmaceutical drugs and for potentially reducing the need for experimental work.

A central problem in drug discovery is the prediction of receptor–ligand binding free energies. Among the many approaches available for free energy calculations, docking and scoring² are among the least computationally expensive but also most approximate. In these methods ligand orientations (poses) are assigned scores, and the quality of the fit is expressed by an empirical function, the scoring function. These scores are used to rank each pose relative to other poses and other ligands. Methods with a higher level of statistical mechanics rigor include molecular mechanics Poisson–Boltzmann surface area (MM-PBSA)³ and molecular mechanics generalized Born surface area (MM-GBSA).⁴ These methods estimate absolute free energies of bound and unbound reference states using molecular dynamics (MD) simulations to sample phase space. Free energies of binding are obtained as averages of interaction energies over snapshots from the MD simulations with entropic contributions calculated from vibrational frequency calculations and the solvation free energy contributions from an implicit solvent model. Although this approach has found extensive usage, especially for the calculation of relative free energies of binding, its accuracy is

limited by the approximate nature of including entropy and solvation effects as well as the force field, which is required to reproduce structures and energies with high accuracy. At the most theoretically rigorous end of the spectrum we have methods, such as potentials of mean force and alchemical free energy calculation approaches.⁵ An example of an alchemical method is thermodynamic integration (TI). It follows an unphysical pathway, where one ligand is mutated to another. It evaluates ratios of partition functions to estimate relative binding free energies and their gradual change during the mutation, which happens in small steps and fully includes the entropic and solvation contributions which are heavily approximated with the less rigorous approaches. In principle, alchemical free energy calculations allow the exact prediction of relative binding free energies, at very high computational cost. However, inadequacies in the force fields used and insufficient sampling introduce errors into the calculated free energies. These errors are exacerbated by ligands that cause changes which are difficult to capture by classical force fields, such as charge transfer and polarization or conformational change on binding, which may require extremely long simulations.

A large number of force fields have been developed and extensively parametrized for common amino acids found in proteins, but the development and parametrization of force fields for general ligands are a much more difficult task. Even in the protein force fields there are issues with their transferability and accuracy, as their form can allow only for an average picture of electronic polarization and no inclusion of electronic charge transfer; yet these effects are ubiquitous and can often make important contributions to energies and structures. Promising progress is being made into polarizable force fields,⁶ but these approaches

Received: December 9, 2010

Published: March 16, 2011

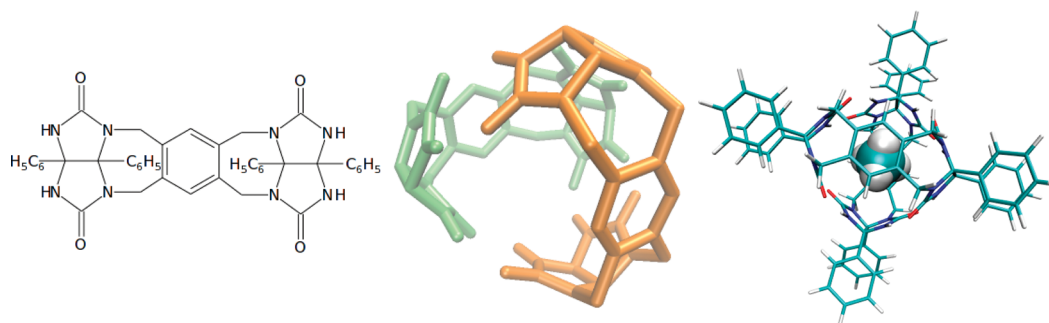


Figure 1. Two-dimensional (2D) diagram of the monomer (left). Truncated structure (2D) of the tennis ball depicting the shape of the cavity (middle). Encapsulation of a methane molecule in the whole dimer (right).

are still not as general or as able to achieve the required levels of accuracy.

First principles quantum mechanical (QM) methods explicitly include the electrons and therefore can fully take into account the electronic charge transfer and polarization and are transferable to any chemical environment. They are therefore ideal for biomolecular simulations, since the interactions between ligand and receptor always involve rearrangement of electrons to a certain extent, and first principles calculations would be expected to provide more accurate binding free energies. The most widely used first principles approach is Kohn–Sham density functional theory (DFT),⁷ as it offers a good compromise between accuracy and computational cost. However, the applicability of first principles calculations in such simulations is limited because in general they are 1000 times more computationally demanding than force field approaches, and more importantly, their computational cost increases with the third power in the number of atoms. In practice this limits the size of the calculations to a few hundred atoms at most, while most biomolecular systems of interest tend to include thousands of atoms. In cases where quantum calculations are unavoidable, such as for example, in the study of chemical reactions in the active sites of enzymes, small parts of the active site are simulated using quantum mechanics while the rest of the system is described by a classical force field. A variety of such QM/MM approaches have been developed,^{8,9} but their application requires extensive experience in order to effect a physically meaningful partition of the system to QM and MM regions and also to properly describe the interaction between these two fundamentally different models. An alternative approach would be to perform first principles calculations on the entire biomolecular system if there was a way to avoid their cubic scaling cost. This can be achieved by using linear-scaling first principles approaches¹⁰ which have the capability for calculations on many thousands of atoms. The development of such approaches has been slow, as it required dealing with a variety of nontrivial physical and computational issues, but today a number of linear-scaling DFT packages are available such as ONETEP,¹¹ CONQUEST,¹² SIESTA,¹³ and others.¹⁴

In this work we are evaluating the use of first principles calculations in combination with a classical force field to simulate host–guest interactions. The system we have selected to study is a model for a protein ligand-binding cavity based on a self-assembling superstructure, the “tennis ball” dimer (Figure 1). We have chosen this model as it combines simplicity with realism and also because there are previous computational¹⁵ studies and experimental¹⁶ data to compare with. We first compare structure optimization with a force field and first principles approaches in

terms of the structural parameters. We then introduce dynamic effects through MD simulations and compare binding energies calculated from MM-PBSA and QM-PBSA to experimental values. For our first principles calculations we use linear-scaling DFT as implemented in the ONETEP¹¹ program which has a demonstrated capability for DFT calculations with thousands of atoms.¹⁷ These are the length scales of several proteins of relevance to current therapeutical challenges, and therefore the use of linear-scaling DFT will allow, with further future testing and validation, the application of first principles-based simulations to some of these proteins.

In Section 2.1 the ONETEP approach is discussed. Section 2.2 will detail the computational methods used in this study. Section 2.3 will outline the procedure and parameters. In Section 3 the results are given and analyzed, and Section 4 summarizes our results and conclusions.

2. METHODS

2.1. The ONETEP Approach. The ONETEP¹¹ program is a linear-scaling DFT code that has been developed for use on parallel computers.¹⁸ ONETEP combines linear scaling with accuracy comparable to conventional cubic-scaling plane-wave methods, which provide an unbiased and systematically improvable approach to DFT calculations. Its novel and highly efficient algorithms allow calculations on systems containing tens of thousands of atoms.¹⁷ ONETEP is based on a reformulation of DFT in terms of the one-particle density matrix. The density matrix in terms of Kohn–Sham orbitals is

$$\rho(\mathbf{r}, \mathbf{r}') = \sum_{n=0}^{\infty} f_n \psi_n(\mathbf{r}) \psi_n^*(\mathbf{r}') \quad (1)$$

where f_n is the occupancy and $\psi_n(\mathbf{r})$ and $\psi_n(\mathbf{r}')$ are the Kohn–Sham orbitals. In ONETEP the density matrix is represented as

$$\rho(\mathbf{r}, \mathbf{r}') = \sum_{\alpha} \sum_{\beta} \phi_{\alpha}(\mathbf{r}) K^{\alpha\beta} \phi_{\beta}^*(\mathbf{r}') \quad (2)$$

where $\phi_{\alpha}(\mathbf{r})$ is the localized nonorthogonal generalized Wannier functions¹⁹ (NGWFs) and $K^{\alpha\beta}$, which is called the density kernel, is the representation of f_n in the duals of these functions. Linear-scaling is achieved by truncation of the density matrix, which decays exponentially for materials with a band gap and by enforcing strict localization of the NGWFs onto atomic regions. In ONETEP as well as optimizing the density kernel, the NGWFs are also optimized, subject to a localization

constraint. Optimizing the NGWFs in situ allows for a minimum number of NGWFs to be used while still achieving plane-wave accuracy. The NGWFs are expanded in a basis set of periodic sinc (psinc) functions,²⁰ which are equivalent to a plane-wave basis as they are related by a unitary transformation. Using a plane-wave basis set allows the accuracy to be improved by changing a single parameter, equivalent to the energy cutoff in conventional plane-wave DFT codes. The psinc basis set provides a uniform description of space, meaning that ONETEP does not suffer from basis set superposition error.²¹

2.2. MM-PBSA and QM-PBSA. The MM-PBSA³ approach is commonly used to calculate relative (and absolute) binding affinities of small molecules to proteins via differences between bound and unbound states. Representative structural ensembles are generated by MD simulations in explicit solvent. Snapshots are extracted at constant time intervals, and the solvent molecules and counterions are removed. Binding energy calculations are then performed on the individual structures using the MM force field together with an implicit solvent approach (PBSA). The free energy of binding is then obtained as an average over the ensemble binding energies. In the solvent model the polar solvent contributions are calculated via the Poisson–Boltzmann (PB) equation, and the nonpolar solvent contributions are calculated from the solvent accessible surface area (SA). The free energy of binding is calculated according to

$$\begin{aligned}\Delta G_{\text{bind}} &= \langle \Delta H_{\text{vac}} \rangle + \langle \Delta G_{\text{solv}} \rangle - T \langle \Delta S \rangle \\ &= \langle \Delta H_{\text{vac}} \rangle + \langle \Delta G_{\text{solv}}^{\text{polar}} \rangle + \langle \Delta G_{\text{solv}}^{\text{nonpolar}} \rangle - T \langle \Delta S \rangle\end{aligned}\quad (3)$$

Where $\langle \Delta H_{\text{vac}} \rangle$ arises from the average difference in van der Waals and electrostatic contributions from the MM force field, $\langle \Delta G_{\text{solv}} \rangle$ is the average free energy of solvation from the PBSA model, and $\langle \Delta S \rangle$ is the entropy of binding which is approximated from harmonic vibrational frequency calculations averaged over the snapshots. A common assumption is that similar ligands bound to the same receptor contribute comparably to the binding entropy, and hence this term is often omitted from the calculations. The difference in free energies of binding between two ligands, A and B, is then given by

$$\Delta \Delta G_{A \rightarrow B} = \Delta \langle \Delta H_{\text{vac}} \rangle_{A \rightarrow B} + \Delta \langle \Delta G_{\text{solv}} \rangle_{A \rightarrow B}\quad (4)$$

The MM-PBSA method used in this study is the single-trajectory approach. In this approach the receptor and ligand structures are taken from the geometry of the complex. It has been observed that this approach produces relative free energies of binding that converge faster with the number of snapshots sampled and are also more accurate, compared to the three-trajectory approach, due to cancellation of errors.²²

A significant source of error in MM-PBSA can be the accuracy of the interaction energies computed for each snapshot, as this depends on the selected force field. Several attempts have been made toward overcoming these limitations by the inclusion of the rigor of quantum mechanics in QM-PBSA extensions of the MM-PBSA approach, which uses semiempirical QM²³ or hybrid QM/MM.^{24,25}

More recently a QM-PBSA approach²⁶ has been presented, where the calculation of the interaction energies by the force field is replaced by DFT calculations on the entire molecules involved. In more detail, the energy of each snapshot is obtained as $E_{\text{QM}} = E_{\text{DFT}} + E_{\text{disp}}$, where E_{disp} is the dispersion correction²⁷ to the

total DFT energy, E_{DFT} . The free energy of solvation for each snapshot from the MM-PBSA calculation is scaled to match the electrostatics of the QM calculation in the following way:

$$\Delta G_{\text{solv}}^{\text{QM}} = \Delta G_{\text{solv}}^{\text{MM}} \left(\frac{\Delta E_{\text{QM}}}{\Delta E_{\text{MM}}} \right)\quad (5)$$

where ΔE_{MM} is the total binding energy from the MM force field, and, as in usual MM-PBSA, is averaged over the snapshots and added to the total DFT energy to give the free energy of binding as

$$\Delta G_{\text{tot}} = \langle \Delta E_{\text{QM}} \rangle + \langle \Delta G_{\text{solv}}^{\text{QM}} \rangle\quad (6)$$

The scaling method used in previous works²⁶ scaled the solvation energy by the fraction of the electrostatic components of the binding energy. In our system that scaling method does not work since dispersion interactions are responsible for most of the binding energy, leading to the MM electrostatic component of the binding energy in the denominator, being very close to zero. We found that the simpler form shown in eq 5 produces reasonable solvation energies.

The first application of QM-PBSA²⁶ has been on protein–protein interactions. The results obtained were in good agreement with the MM-PBSA, most likely because the force field employed has been extensively and carefully parametrized for protein systems and improved over a number of years. As our present system does not consist of amino acids, we do not have the advantage of using such a well-developed force field. This is a situation which is common in drug design as nonstandard residues and new ligands are explored and the reliability of a general force field needs to be checked on a case by case basis. Here we are aiming to investigate how QM-PBSA can be used in such a case, as an accuracy benchmark for MM-PBSA or as an alternative approach.

2.3. Simulation Details. The tennis ball structure was built and loosely minimized with the MOE²⁸ program. MM simulations were carried out using the AMBER 10²⁹ package. The tennis ball was modeled using the generalized AMBER force field³⁰ (gaff) and solvated with the CHCl₃ explicit solvent model (as implemented in AMBER 10) in a periodic box.

To equilibrate the system, the hydrogens were relaxed keeping all heavy atoms restrained in the host and solvent, then relaxing the solvent with restraints still on the host. The system was heated to 300 K still restraining the host for 200 ps with the NVT ensemble and ran for a further 200 ps with the NPT ensemble at 300 K in order to equilibrate the solvent density. This was cooled to 100 K over 100 ps and minimization's carried out reducing the restraints on the host heavy atoms in stages (500, 100, 50, 20, 10, 5, 2, 1, and 0.5 kcal mol⁻¹ Å⁻²). Finally the system was heated to 300 K with no restraints over 200 ps and then ran for a further 200 ps at 300 K with the NPT ensemble, at the end of which the root mean squared deviation of the carbon, nitrogen, and oxygen atoms was converged and less than 0.8 Å relative to the starting frame. production simulations were run for 2 ns with the NPT ensemble at 300 K. All MD simulations used the Langevin thermostat, the particle mesh Ewald (PME) sum for the electrostatic interactions, a time-step of 2 fs, and the SHAKE algorithm.³¹ For the MM-PBSA calculation an infinite non-bonded cutoff was used with a dielectric constant of 4.5 to represent the chloroform solvent. All ONETEP single-point energies were converged to 0.0002 hartree (~ 0.1 kcal mol⁻¹). Four NGWFs were used to describe carbon, oxygen and

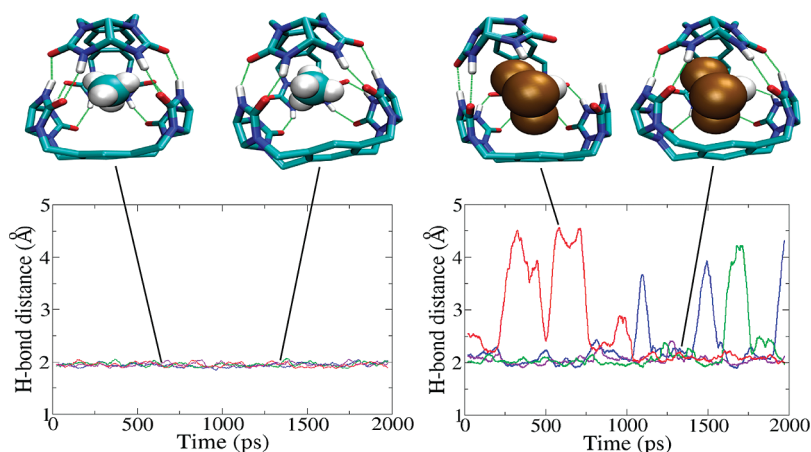


Figure 2. Plots of the hydrogen-bond lengths from four hydrogen-bonding positions (C=O's of top monomer to H—N's of bottom monomer) in the CH₄ complex (left) and CHCl₃ complex (right). Structures taken as “snapshots” at two points of the simulations are shown, the green dashed lines represent the hydrogen bonds present at each snapshot. In the graphs, the four colored lines correspond to the four hydrogen bonds measured in each complex (phenyl rings not displayed).

nitrogen, 1 for hydrogens and 9 NGWFs for the halogen atoms. A kinetic energy cutoff of 800 eV for the psinc basis set was used, with the GGA exchange–correlation functional PBE³² combined with our implementation of the DFT+D approach to account for dispersion parametrized specifically for this functional.²⁷

3. RESULTS AND DISCUSSION

3.1. Validation tests. In cases where two different approaches are used to explore the conformational space, the compatibility of the methods used is an important consideration.³³ First, it is desirable that the minima on the potential energy surface between the QM and the MM approach are as close as possible. To investigate this we have carried out geometry optimizations of the three complexes using ONETEP and AMBER. We have also carried out further validation of the QM approach by doing the same geometry optimizations with the Gaussian³⁴ program which can perform all-electron DFT calculations with Gaussian basis sets. For these all-electron calculations we used a correlation consistent split valence basis set (cc-pVDZ³⁵) and the B97 exchange–correlation functional³⁶ with the DFT+D approach for including dispersion contributions, as parametrized by Grimme et al.³⁷ The structural parameters between the optimized geometries by the three methods were compared. Bond lengths vary by less than 0.03 Å, and internal angles, such as those within rings, vary by less than 0.5, with the more flexible angles differing by 2–3. Hydrogen bonds from ONETEP (Gaussian) are shorter than these from the AMBER optimized structure by 0.2 Å (0.1 Å), and the distance separating the monomers differs by as much as 0.5 Å between the ONETEP and AMBER structures. All the methods predict hydrogen bonds which are longer by around 0.2 Å for the CHCl₃ complex compared to the tennis ball complexed with CH₄ or CF₄ and the empty dimer, which is to be expected as the CHCl₃ is slightly larger than the size of the empty cavity.

As we are interested in properties at finite temperatures (usually room temperature), using only equilibrium geometries is not sufficient, as dynamical motion causes the molecules to visit many configurations which can differ from the relaxed structures. Thus, MD simulations are run for time scales which are long enough (ns) to sample the dynamical behavior of this

Table 1. Average (Maximum) of Forces on Atoms from AMBER and ONETEP from 10 Snapshots^a

complex	ONETEP	AMBER
CH ₄	29.3 (153.0)	29.9 (107.0)
CHCl ₃	29.3 (147.8)	29.7 (105.0)
CF ₄	30.1 (150.8)	30.1 (128.1)

^a Values in kcal/mol/Å.

system, using the classical force field approach. The importance of accounting for dynamic motion for the tennis ball system is shown in Figure 2. Here we examine hydrogen-bond lengths in the CH₄ and CHCl₃ complexes throughout the 2 ns MD simulations. During the simulation, the hydrogen bonds in the CH₄ complex are stable, staying at around 2 Å. In contrast, the hydrogen bonds in the CHCl₃ complex are intermittent. We observe that the dimer opens at a point, around 4 Å, then moves back into position, re-establishes the hydrogen bond, and breaks at another point. This happens due to the size of the chloroform ligand; it is too large to fit comfortably between the monomers causing the cavity to open and close during the simulation. Figure 2 demonstrates that the CHCl₃ complex has one hydrogen-bond broken most of the time. In this case the minimum energy structure, which has all the hydrogen bonds intact, albeit elongated, will not provide an adequate description of the ensemble of structures encountered at room temperature. We can demonstrate this further by noting that the binding energy for the CHCl₃ complex as calculated with ONETEP on the optimized structure is 2.6 kcal mol⁻¹, while when taking into account 200 snapshots extracted from the MD ensemble, it is -7.1 kcal mol⁻¹ (see Section 3.2), in close agreement with the experimental value of -7 kcal mol⁻¹. As a dynamical ensemble of structures is necessary for this study, we also need to confirm that the conformations sampled by the force field are not unphysical as far as the QM potential energy surface is concerned. To explore this issue, we have compared forces on atoms calculated from ONETEP and AMBER on several of the snapshots. An indication that the compatibility of the two approaches is good in this case is given by the values reported in Table 1, which presents the average (maximum) of the absolute values of the force on all

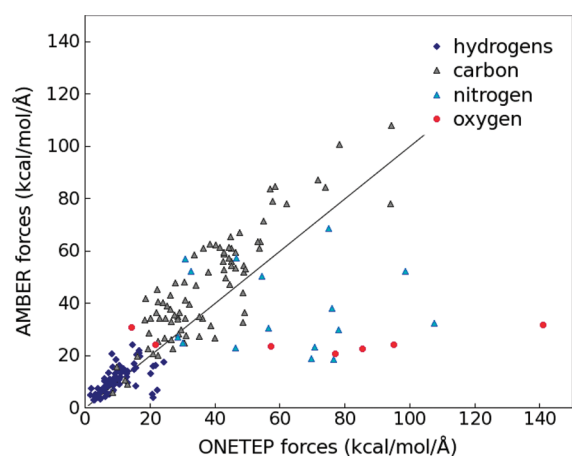


Figure 3. Correlation between $|F_{\text{QM}}|$ and $|F_{\text{MM}}|$ for a single snapshot of the CH_4 complex. Other snapshots and complexes show similar behavior.

Table 2. MM- and QM-PBSA Results Presenting Binding Free Energies Separated into Differences in Enthalpy and Free Energy and Relative Differences between Ligands Using CH_4 as a Reference^a

	ΔH			ΔG		
	MM-PBSA	QM-PBSA	expt ¹⁶	MM-PBSA	QM-PBSA	expt ¹⁶
CH_4	-8.7 ± 0.05	-9.0 ± 0.04	-9	-8.5 ± 0.05	-8.9 ± 0.05	-3.0
CHCl_3	-16.8 ± 0.19	-7.1 ± 0.27	-7	-14.9 ± 0.16	-6.2 ± 0.23	2.2
CF_4	-12.6 ± 0.09	-6.0 ± 0.08	N/A	-11.9 ± 0.08	-5.7 ± 0.08	-0.2^{15}

	$\Delta\Delta H$			$\Delta\Delta G$		
	MM-PBSA	QM-PBSA	expt ¹⁶	MM-PBSA	QM-PBSA	expt ¹⁶
$\text{CH}_4 \rightarrow \text{CHCl}_3$	-8.1	2.1	2	-6.4	2.7	5.2
$\text{CH}_4 \rightarrow \text{CF}_4$	-3.9	3.0	N/A	-3.4	3.2	2.8^{15}

^a ΔH is the energy in vacuum. ΔG is the vacuum energy plus the solvation energy (for MM- and QM-PBSA this term does not include conformational entropy).

atoms over 10 equally spaced snapshots. Even though these agree extremely well between the two approaches, if we look in more detail at individual atoms, the agreement is not so good. The largest difference between the QM and MM forces on any single atom is $80 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$, but for most atoms, it is less than $20 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$. Figure 3 compares the forces between QM and MM between individual atoms for a single snapshot of the CH_4 complex, colored according to the type of element. We can observe that for hydrogen and carbon atoms both ONETEP and AMBER forces agree reasonably well. The large differences on the oxygen and nitrogen atoms are as expected, this is because the parametrization of the ligand is done in group-wise fashion, so an amine-group will have a charge of one, but it is not clear how the charges are distributed over the atoms, thus the charges for heteroatoms will strongly differ from ONETEP causing a strong difference in gradient. This behavior is representative of all snapshots for the three systems. Our comparisons show that there is substantial variability in the forces obtained with the two approaches, however the forces in both cases are within expected ranges, and the average forces are comparable. This suggests that no unphysical conformations are visited by the force field.

Having established the importance of taking into account the dynamical behavior of this system, we finally tested the

Table 3. Relative Binding Free Energies (kcal/mol) Obtained via TI by Fox et al.^a

	TI ¹⁵	TI with gaff	QM-PBSA	MM-PBSA	expt ¹⁶
$\text{CH}_4 \rightarrow \text{CHCl}_3$	7.8	7.2	2.8	-6.4	5.2
$\text{CH}_4 \rightarrow \text{CF}_4$	0.9	0.1	3.2	-3.4	2.8

^a TI results using the gaff and results from our QM-PBSA approach.

convergence of PBSA energies as a function of the number of MD snapshots. An increasing number of snapshots was used, obtained by sampling uniformly through the 2 ns production simulations. From each simulation, 50, 100, 160, and 200 equally spaced snapshots were extracted to study the convergence. We found that the variation in the binding free energies calculated in ONETEP or AMBER when going from 50 to 200 snapshots was less than 0.2 kcal/mol for all systems studied. All MM- and QM-PBSA results we report here were obtained using 200 snapshots.

3.2. Free Energies of Binding. The energies of binding that were obtained with the MM- and QM-PBSA approaches for all the complexes are presented in Table 2. The table shows the enthalpies of binding (ΔH) computed from either the force field or the DFT calculations with ONETEP and the free energy of binding (ΔG), which includes solvation contributions. We observe that for CH_4 AMBER predicts a ΔH that agrees well with experiment (to $<0.3 \text{ kcal mol}^{-1}$), however it overestimates the ΔH for the halogen-containing ligands to over twice the experimental value. This suggests that the force field does not capture well the interaction energies of the halogen atoms with the cavity. ONETEP produces ΔH values that are in close agreement (within $0.1 \text{ kcal mol}^{-1}$) to the experimentally determined ΔH values, which supports further our earliest observation that the ensemble of structures provided by the force field has a high overlap with the QM ensemble. The larger standard errors in the calculated energy differences for the CHCl_3 complex, $0.27 \text{ kcal mol}^{-1}$ compared to $0.04 \text{ kcal mol}^{-1}$ for CH_4 , are expected since this structure shows considerably more fluctuation than the other complexes, as we saw in Figure 2.

Since AMBER overestimates the interaction energies for the halogen-containing ligands, the calculated $\Delta\Delta G$'s predict a more favorable interaction than was found experimentally and in the previous computational study. Our improvements by the QM calculations refer to the enthalpic part of the binding energies, and indeed we can observe that the $\Delta\Delta H$ values are a very good match to experiment.

As the enthalpy is accounted for so well, and the free energy of solvation in this case makes a minimal contribution due to the nonaqueous solvent, the large discrepancy in the free energy differences ΔG can be attributed to the neglect of configurational entropy. When considering the $\Delta\Delta G$ values a large fraction of this error is canceled, and we obtain reasonably good agreement with experiment [2.7 versus $5.2 \text{ kcal mol}^{-1}$ for $\Delta\Delta G(\text{CH}_4 \rightarrow \text{CHCl}_3)$ and 3.2 versus $2.8 \text{ kcal mol}^{-1}$ for $\Delta\Delta G(\text{CH}_4 \rightarrow \text{CF}_4)$] for ONETEP, while the AMBER values show discrepancies of more than 5 kcal mol^{-1} , precisely due to the bad estimation of enthalpy.

Previous computational results by Fox et al.¹⁵ were obtained from TI calculations. Simulations were performed with AMBER 4.1 using the all-atom force field by Cornell et al.³⁸ and the partial charges obtained from a multiple molecule RESP fit. Table 3 compares their results to TI free energies we obtained with AMBER 10 using the gaff force field and with our MM- and QM-

PBSA results. We observe that both TI approaches obtain comparable relative binding free energies (7.8 versus 7.2 kcal mol⁻¹ for CH₄ → CHCl₃ and 0.9 versus 0.1 kcal mol⁻¹ for CH₄ → CF₄) and considerably better than MM-PBSA ($\Delta\Delta G(\text{CH}_4 \rightarrow \text{CHCl}_3)$ of 7.8 kcal mol⁻¹ rather than -6.4 kcal mol⁻¹). While TI is a more rigorous approach which fully accounts for entropic effects, we observe that our QM-PBSA energies achieve improved agreement with experiment. So at least in this system the accurate description of interaction energies that is provided by the DFT calculations is critical for the correct calculation of free energy differences.

As we have mentioned in the Introduction, force fields are significantly less computationally demanding than first principles quantum calculations, and this is reflected in our timings. For example, a single-point energy force field calculation on one of our complexes takes about 0.35 core seconds on an Intel CORE2 machine, while the same calculation with DFT takes about 24 core hours on the same computational platform. Therefore in terms of throughput, the force field calculations have a clear advantage. However, the point is that in several cases the unbiased and accurate description that is provided by the first principles calculations can be indispensable. For example, electronic polarization, or halogen- π interactions, which are poorly described by available force fields. We therefore expect that large-scale first principles quantum calculations will be a valuable tool in the final stages of computational drug design where careful refinement is required. The linear-scaling formalism makes it feasible to extend the application of these calculations to biomolecules with thousands of atoms, especially in combination with new HPC technologies such as GPUs and peta-scale supercomputers.

4. CONCLUSIONS

In this paper we have presented an approach for reducing some of the limitations of the MM-PBSA method. Toward this aim we have used the ONETEP program to calculate the QM interaction energies with solvation contributions extracted from a traditional MM-PBSA method and scaled to match the QM energies. Conformational space was sampled with classical force field molecular dynamics simulations, and the compatibility of the structural ensemble, with respect to the potential energy surface, was checked by comparing forces on atoms between the two methods. These showed that, although there was substantial variation in the forces obtained with the two approaches, the forces in both cases were within expected ranges and that no unphysical conformations appear to be visited by the force field. This QM-PBSA approach obtained energies which are significantly improved over the MM computed energies, with enthalpic energies agreeing with experimental ΔH values to within 0.1 kcal mol⁻¹. The neglect of entropy leads to poor agreement with experimental absolute binding free energy values, however, relative binding free energies show considerable improvement agreeing well with experiment. These even show an improvement over the more rigorous TI method.

While the model we have studied is relatively simple and small (for biomolecular standards), it does include important and difficult to capture interactions, such as halogen- π interactions, which are not at all well described by force fields and even hydrogen bonds whose accurate description by nonquantum methods is reasonable but cannot be taken for granted. Therefore this is a small but important step toward modeling some of the

crucial interactions in real protein-ligand systems. Armed with the experience from this study and with the ability of ONETEP for DFT calculations with thousands of atoms, we can in the future extend our investigation with QM-based free energy approaches to protein-ligand complexes of pharmaceutical interest. These typically include further challenges, such as rotatable bonds in the ligand, ligand and pocket desolvation, and partial solvation of the bound ligand in the case of solvent-exposed binding pockets.

AUTHOR INFORMATION

Corresponding Author

*E-mail: c.skylaris@soton.ac.uk.

ACKNOWLEDGMENT

S.F. would like to thank BBSRC for a CASE studentship award supported by Boehringer Ingelheim. C.-K.S. would like to thank the Royal Society for a University Research Fellowship. The calculations in this work were carried out on the Iridis3 Supercomputer of the University of Southampton. We would like to thank Dr. Danny Cole from the Theory of Condensed Matter group at the University of Cambridge for useful discussions.

REFERENCES

- (1) Friesner, R. A.; Dunietz, B. D. *Acc. Chem. Res.* **2001**, *34*, 351–358.
- (2) Halperin, I.; Ma, B. Y.; Wolfson, H.; Nussinov, R. *Proteins: Struct., Funct., Genet.* **2002**, *47*, 409–443.
- (3) 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.
- (4) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.
- (5) Michel, J.; Essex, J. J. *Comput.-Aided. Mol. Des.* **2010**, *24*, 639–658.
- (6) Halgren, T. A.; Damm, W. *Curr. Opin. Struct. Biol.* **2001**, *11*, 236–242.
- (7) Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *140*, A1133–A1138.
- (8) Kamerlin, S. C. L.; Haranczyk, M.; Warshel, A. *J. Phys. Chem. B* **2009**, *113*, 1253–1272.
- (9) Riccardi, D.; Schaefer, P.; Yang, Y.; Yu, H. B.; Ghosh, N.; Prat-Resina, X.; Konig, P.; Li, G. H.; Xu, D. G.; Guo, H.; Elstner, M.; Cui, Q. *J. Phys. Chem. B* **2006**, *110*, 6458–6469.
- (10) Goedecker, S. *Rev. Mod. Phys.* **1999**, *71*, 1085–1123.
- (11) Skylaris, C.-K.; Haynes, P. D.; Mostofi, A. A.; Payne, M. C. *J. Chem. Phys.* **2005**, *122*, 084119.
- (12) Bowler, D. R.; Bush, I. J.; Gillan, M. J. *J. Quant. Chem.* **2000**, *77*, 831–842.
- (13) Soler, J. M.; Artacho, E.; Gale, J. D.; Garcia, A.; Junquera, J.; Ordejon, P.; Sanchez, D. *J. Phys.: Condens. Matter* **2002**, *14*, 2745–2779.
- (14) Bowler, D. R.; Fattbert, J. L.; Gillan, M. J.; Haynes, P. D.; Skylaris, C.-K. *J. Phys.: Condens. Matter* **2008**, *20*, 290301.
- (15) Fox, T.; Thomas, B. E., IV; McCarrick, M.; Kollman, P. A. *J. Phys. Chem.* **1996**, *100*, 10779–10783.
- (16) Branda, N.; Wyler, R.; Rebek, J. *Science* **1994**, *263*, 1267–1268.
- (17) Hine, N. D. M.; Haynes, P. D.; Mostofi, A. A.; Skylaris, C.-K.; Payne, M. C. *Comput. Phys. Commun.* **2009**, *180*, 1041–1053.
- (18) Skylaris, C.-K.; Haynes, P. D.; Mostofi, A. A.; Payne, M. C. *Phys. Status Solidi* **2006**, *243*, 973–988.
- (19) Skylaris, C.-K.; Mostofi, A. A.; Haynes, P. D.; Diéguez, O.; Payne, M. C. *Phys. Rev. B* **2002**, *66*, 035119.
- (20) Mostofi, A. A.; Haynes, P. D.; Skylaris, C.-K.; Payne, M. C. *J. Chem. Phys.* **2003**, *119*, 8842–8848.

- (21) Haynes, P. D.; Skylaris, C.-K.; Mostofi, A. A.; Payne, M. C. *Chem. Phys. Lett.* **2006**, *422*, 345–349.
- (22) Massova, I.; Kollman, P. A. *J. Am. Chem. Soc.* **1999**, *121*, 8133–8143.
- (23) Diaz, N.; Suárez, D.; Merz, K. M.; Sordo, T. L. *J. Med. Chem.* **2005**, *48*, 780–791.
- (24) Kaukonen, M.; Soderhjelm, P.; Heimdal, J.; Ryde, U. *J. Phys. Chem. B* **2008**, *112*, 12537–12548.
- (25) Wang, M.; Wong, C. F. *J. Chem. Phys.* **2007**, *126*, 026101.
- (26) Cole, D. J.; Skylaris, C.-K.; Rajendra, E.; Venkitaraman, A. R.; Payne, M. C. *Europhys. Lett.* **2010**, *91*, 37004.
- (27) Hill, Q.; Skylaris, C.-K. *Proc. R. Soc. A* **2009**, *465*, 669–683.
- (28) MOE, 2009.10; Chemical Computing Group, Inc.: Montreal, Canada, 2009.
- (29) Case, D. A.; Darden, T.; Cheatham, T.; Simmerling, C.; Wang, J.; Duke, R.; Luo, R.; Crowley, M.; Roitberg, S. H. A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Matthews, D.; Seetin, M.; C.; Sagui, Babin, V.; Kollman, P. A. *AMBER10*; University of California, San Francisco: San Francisco, CA, 2008.
- (30) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157–1174.
- (31) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.
- (32) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- (33) Robinson, M.; Haynes, P. D. *J. Chem. Phys.* **2010**, *133*, 084109.
- (34) 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. J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, revision A.1. Gaussian, Inc.: Wallingford, CT, 2009.
- (35) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007–1023.
- (36) Becke, A. D. *J. Chem. Phys.* **1997**, *107*, 8554–8560.
- (37) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787–1799.
- (38) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Fergusson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

Wavelet Transform for Spectroscopic Analysis: Application to Diols in Water

Francesco Muniz-Miranda,[†] Marco Pagliai,[‡] Gianni Cardini,^{*,†,‡} and Vincenzo Schettino^{†,‡}

[†]European Laboratory for Non-Linear Spectroscopy (LENS), via Nello Carrara 1, 50019, Sesto Fiorentino (FI), Italy

[‡]Dipartimento di Chimica "Ugo Schiff", Università degli Studi di Firenze, via della Lastruccia 3, 50019, Sesto Fiorentino (FI), Italy

ABSTRACT: Wavelet transform has been used to correlate spectroscopic and structural properties from trajectories obtained by *ab initio* molecular dynamics simulations. This method has been applied to hydrogen bond dynamics of glycols in heavy water solutions, showing how the stretching frequency of the *intramolecular* O–H bond changes with the *intermolecular* hydrogen-bond distance. The resulting wavelet spectrograms have been interpreted according to H-bond strength and stability.

1. INTRODUCTION

Classical and *ab initio* molecular dynamics (MD) simulations are powerful tools used to study the structure and dynamics of condensed systems. In particular, the Car–Parrinello approach is well suited to analyze condensed systems with fairly strong intermolecular interactions, like hydrogen bonding. For these specific interactions, the vibrational density of states (VDOS) and the vibrational spectra are particularly significant probes of the structural and dynamic properties of the system. Fourier Transform (FT) is the most commonly used mathematical tool to extract spectral information from time-resolved signals and from trajectories of molecular dynamics simulations. However, there is additional information hidden in the MD trajectories that cannot be sorted out by this approach. In fact, FT extracts the frequency content of a time signal, but it is unable to localize the frequency of a signal in time. This information could be relevant in enlightening the dynamics of the system and in the deconvolution of the steady state spectroscopic response. The wavelet transform (WT) is another type of integral transform, which is able to perform both time-frequency and multiresolution analysis, and for this reason, its use in theoretical and computational chemistry is gradually increasing. In particular, it has been applied to electronic structure calculations^{1–5} and to molecular dynamics^{6–8} also to obtain time-dependent vibrational frequencies.^{9,10}

In the present work, WT has been used to correlate vibrational frequencies of two molecular systems with other time-dependent structural properties. The simulated molecular systems are two glycols in heavy water solutions, which have been investigated by *ab initio* molecular dynamics, using the Car–Parrinello method,¹¹ to take into account polarization effects and charge transfer and correctly describe the hydrogen-bond interactions. Therefore, structural and dynamic properties of the diols have been analyzed from the trajectories of the simulations.

WT has been employed to improve the analysis of H-bond dynamics and its relation with spectroscopic properties. Since in molecular dynamics simulations it is always possible to associate to each simulation step a structural quantity, a distance–frequency

correlation showing how the VDOS changes with a particular H-bond length has been obtained.

The glycols simulated in the present work are propane-1,3-diol (propanediol, PDO) and ethane-1,2-diol (ethylene glycol, EG), two homologous compounds of industrial interest that interact with the aqueous solvent mainly through *intermolecular* hydrogen bonds. PDO is a transparent, nontoxic liquid glycol that can be obtained by fermentation of sugars and can replace other glycols in formulations where petroleum-free ingredients are needed.¹² EG is the lower homologue of PDO and is mainly used as an antifreeze and a medium for convective heat transfer due to its low freezing point. For both PDO and EG, a deep knowledge of the hydrogen-bond interactions and of the solvation dynamics is needed to improve their use in industrial applications. These glycols are study cases of particular interest since they have two interaction sites with the water solvent, and the analysis of the hydrogen bond features is challenging because of the overlap of the solvent–solvent and solute–solvent contributions.

The structure of the paper is as follows. In section 2, details of the computational procedure are reported, and the essentials of the wavelet analysis are described together with the implementation carried out in the present work. In section 3, the structural and electronic properties obtained with the usual Fourier Transform analysis are described, and then, the time-frequency analysis and the structure-frequency correlation obtained by wavelet analysis are discussed. Section 4 contains some concluding remarks.

2. COMPUTATIONAL DETAILS

Ab initio molecular dynamics simulations have been performed for PDO and EG in water. These simulations have been carried out with the CPMD package¹³ in the microcanonical ensemble (NVE) in conjunction with the BLYP^{14,15} exchange and correlation functional, in cubic boxes with edges of 12.7005 Å (for PDO) and 12.6819 Å (for EG), with periodic boundary conditions. The samples in the two simulation boxes are made of

Received: November 2, 2010

Published: February 23, 2011

Table 1. Structural Parameters

ethane-1,2-diol ^a	Howard et al. ²²	BLYP ^b	B3LYP ^b	MP2 ^b	CPMD ^c
$r(\text{C}-\text{C})$	1.514	1.527	1.517	1.513	1.523
$r(\text{C}-\text{O}_1)$	1.433	1.454	1.434	1.431	1.457
$r(\text{C}-\text{O}_2)$	1.421	1.440	1.421	1.418	1.444
$r(\text{O}_1-\text{H})$	0.961	0.971	0.961	0.961	0.974
$r(\text{O}_2-\text{H})$	0.964	0.975	0.965	0.964	0.977
$r(\text{O}_1\cdots\text{H})$	2.331	2.434	2.399	2.323	2.418
$\theta(\text{O}_1-\text{C}-\text{C})$	106.1	106.7	106.8	106.2	106.8
$\theta(\text{O}_2-\text{C}-\text{C})$	111.2	112.2	112.0	111.1	112.2
$\phi(\text{H}-\text{O}_1-\text{C}-\text{C})$	-166.0	-164.8	-166.7	-164.0	-166.6
$\phi(\text{H}-\text{O}_2-\text{C}-\text{C})$	-51.5	-54.9	-53.6	-51.9	-52.1
propane-1,3-diol ^a	Kinnegeing et al. ²³	BLYP ^b	B3LYP ^b	MP2 ^b	CPMD ^c
$r(\text{C}_3-\text{C}_4)$	1.514	1.533	1.523	1.518	1.528
$r(\text{C}_4-\text{C}_5)$	1.514	1.543	1.532	1.527	1.539
$r(\text{C}_3-\text{O}_1)$	1.410	1.461	1.439	1.435	1.465
$r(\text{C}_5-\text{O}_2)$	1.410	1.441	1.422	1.420	1.445
$r(\text{O}_1-\text{H})$	1.040	0.972	0.962	0.962	0.975
$r(\text{O}_2-\text{H})$	0.980	0.977	0.967	0.965	0.979
$r(\text{O}_1\cdots\text{H})$	1.753	2.051	2.034	2.000	2.052
$\theta(\text{O}_1-\text{C}_3-\text{C}_4)$	112.0	108.4	108.6	108.0	108.9
$\theta(\text{O}_2-\text{C}_5-\text{C}_4)$	108.0	113.4	113.4	113.0	114.0
$\theta(\text{C}_3-\text{C}_4-\text{C}_5)$	112.0	113.9	113.8	112.9	114.1
$\phi(\text{H}-\text{O}_1-\text{C}_3-\text{C}_4)$	180.0	173.9	175.6	173.2	172.3
$\phi(\text{H}-\text{O}_2-\text{C}_5-\text{C}_4)$	-46.0	-46.3	-46.7	-48.4	-41.4

^a Bond lengths in Å, angles in degrees. ^b DFT and MP2 calculations have been performed with the GAMESS¹⁹ suite of programs in conjunction with the 6-311++G(d,p) basis set. ^c CPMD geometry optimizations have been performed with the same parameters reported for the molecular dynamics simulations.

one solute molecule and 64 D₂O molecules (at the experimental density of deuterated water, $\sim 1.106 \text{ g cm}^{-3}$). Norm-conserving Martins–Troullier¹⁶ pseudopotentials have been used along with Kleinman–Bylander¹⁷ decomposition, and the plane waves expansions have been truncated at 85 Ry (this choice has been shown to be particularly effective in CPMD simulations¹⁸). In order to use a larger time step (δt of 5 au ~ 0.12 fs), hydrogen atoms have been replaced by deuterium atoms. The fictitious electronic mass has been set at 700 au to allow for a good decoupling between electronic and nuclear degrees of freedom. The systems have been thermalized by scaling the atomic velocities for 2 ps in order to keep the temperature at the value of 300 ± 50 K, and the trajectories for both systems have been collected for ~ 32 ps.

The data collected from the two simulations of the glycols have been compared with corresponding CPMD simulations in vacuum. The computational parameters used in vacuum simulations, such as cell parameters, energy cutoff, and time step, are the same as those in water simulations. To further check the accuracy in the reproduction of the structural parameters for the isolated diols with the computational approach adopted, geometry optimizations have been carried out with the GAMESS package¹⁹ at the MP2 and density functional theory (DFT) levels of theory (using the BLYP^{14,15} and B3LYP^{15,20,21} exchange-correlation functionals) with the 6-311++G(d,p) basis set.

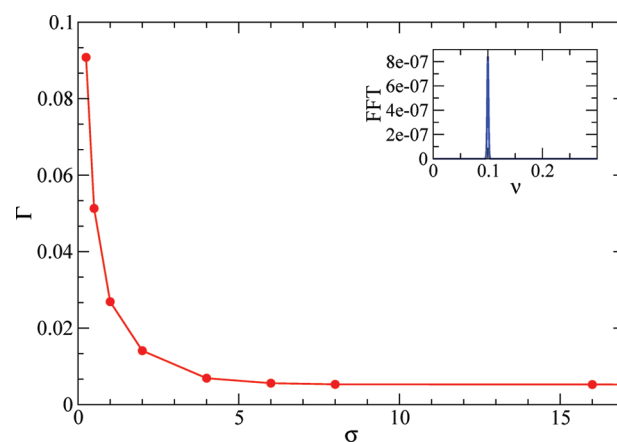


Figure 1. Spread of the wavelet spectrum of time signal in eq 8 as a function of the σ parameter (red). In the inset, the fast-Fourier transform (FFT) is reported as a blue line.

A comparison of experimental and calculated data is reported in Table 1, showing only minor differences between all electrons and pseudopotential calculations. All calculated normal mode frequencies are real and positive, ensuring that the systems are in their respective equilibrium configurations.

2.1. Wavelet Analysis. The FT is the conventional method used to extract the frequency content of a time-resolved signal $f(t)$:

$$\hat{f}(\omega) = \int_{-\infty}^{+\infty} dt e^{-i\omega \cdot t} f(t) \quad (1)$$

A limitation of such an integral transform is that it cannot show how the frequency of interest changes with time due to variation of the intermolecular interactions.

Recently, it has been shown that WT can be applied to chemical problems,^{24,25} and particularly to molecular dynamics simulations in order to sort out both structural and time-dependent properties.^{6–10} Actually, WT can be seen as a mathematical tool that performs time-frequency analysis by using a set of time-window functions called wavelets, $\psi_{n,s}$, that sample the $f(t)$ signal. The $\psi_{n,s}$ wavelets are obtained by time-translation and time-dilatation of a generating function called “mother wavelet”. These features are regulated by the couple of n and s integer parameters, as will be discussed in more detail in the following. In the present work, the Morlet-Gabor mother wavelet²⁶ has been used, which is a Gaussian-like function multiplied by a plane wave defined as

$$\psi(t) = \frac{1}{\pi^{1/4}} e^{i\omega_0 \cdot t} e^{-t^2/2\sigma^2} \quad (2)$$

where ω_0 and σ represent the main oscillation frequency of the plane wave and the width at half-height of the Gaussian time-window, respectively. Kirby²⁷ showed that this mother wavelet is one of the most successful in reproducing the Fourier power spectra.

This function is translated and stretched by the n and s parameters, to give the entire set of wavelet functions $\{\psi_{n,s}\}$:

$$\psi_{n,s}(t') = \psi\left(\frac{t' - n \cdot t}{s}\right) \quad (3)$$

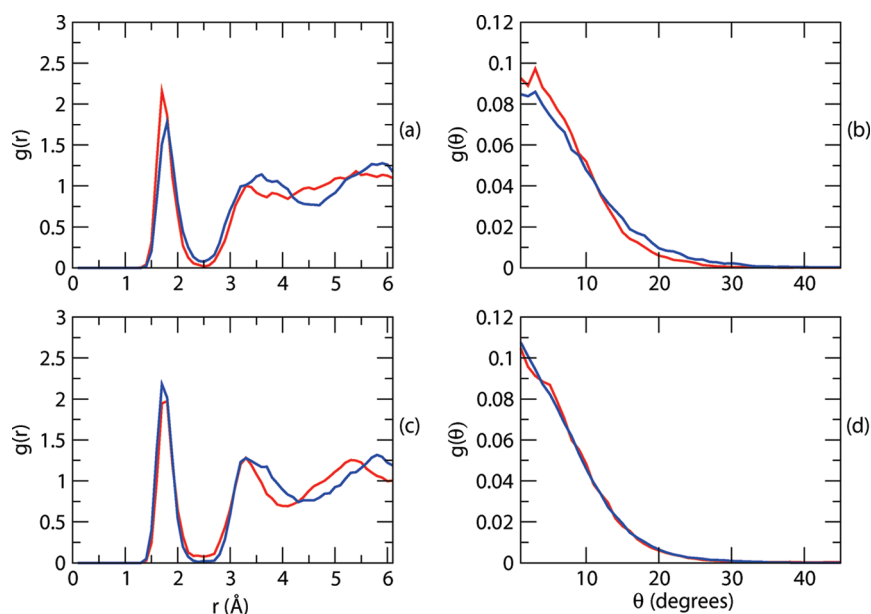


Figure 2. Accepter $g(r)$ and $g(\theta)$: red for “site 1”, blue for “site 2”; panels a and b for EG, panels c and d for PDO.

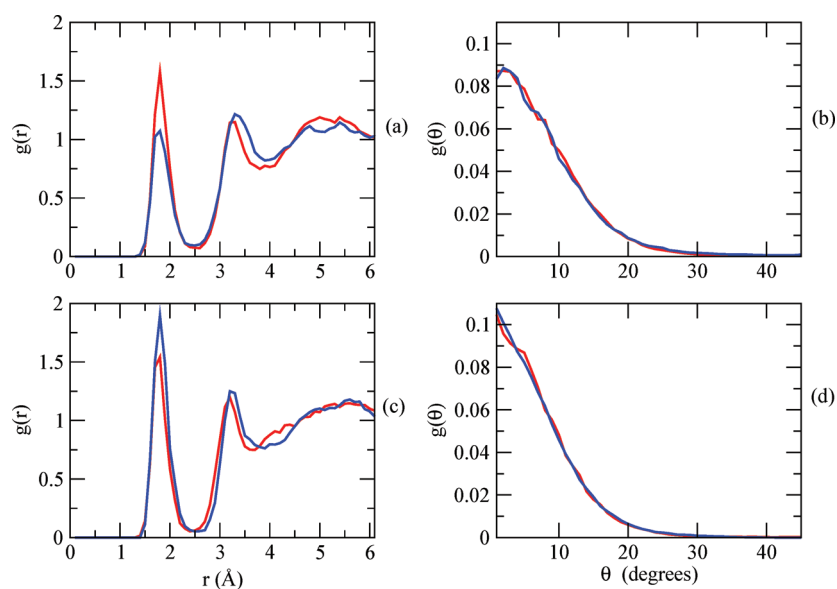


Figure 3. Donor $g(r)$ and $g(\theta)$: red for “site 1”, blue for “site 2”; panels a and b for EG, panels c and d for PDO.

Table 2. Average Coordination Numbers

	site 1	site 2
EG donor	1.00	1.01
EG acceptor	1.70	1.42
PDO donor	0.98	1.00
PDO acceptor	1.56	1.94

The n parameter localizes the frequency in time, and $1/s$ is proportional to the frequency, whose detailed meaning will be illustrated in the following. The continuous WT is given by

$$\mathcal{W}(n, s) = \int_{-\infty}^{+\infty} dt' f(t') \psi_{n,s}^*(t') \quad (4)$$

while the discretized expression used in this work is²⁸

$$\mathcal{W}(n, s) = \sum_{n'=0}^{N-1} f(n' \cdot \delta t) \psi^* \left[\frac{(n' - n) \cdot \delta t}{s} \right] \quad (5)$$

In eq 5, the product $n' \delta t$ represents the total time at the n' th time-step of the molecular dynamics and localizes the signal in time. Thus, the wavelet transform $\mathcal{W}(n, s)$ gives the frequency content of the signal $f(t)$ over a Gaussian time-window centered at $n \delta t$. For the Morlet-Gabor set of basis functions,²⁸ the wavelength is defined as

$$\lambda = \frac{s4\pi}{\omega_0 + \sqrt{2 + \omega_0^2}} \quad (6)$$

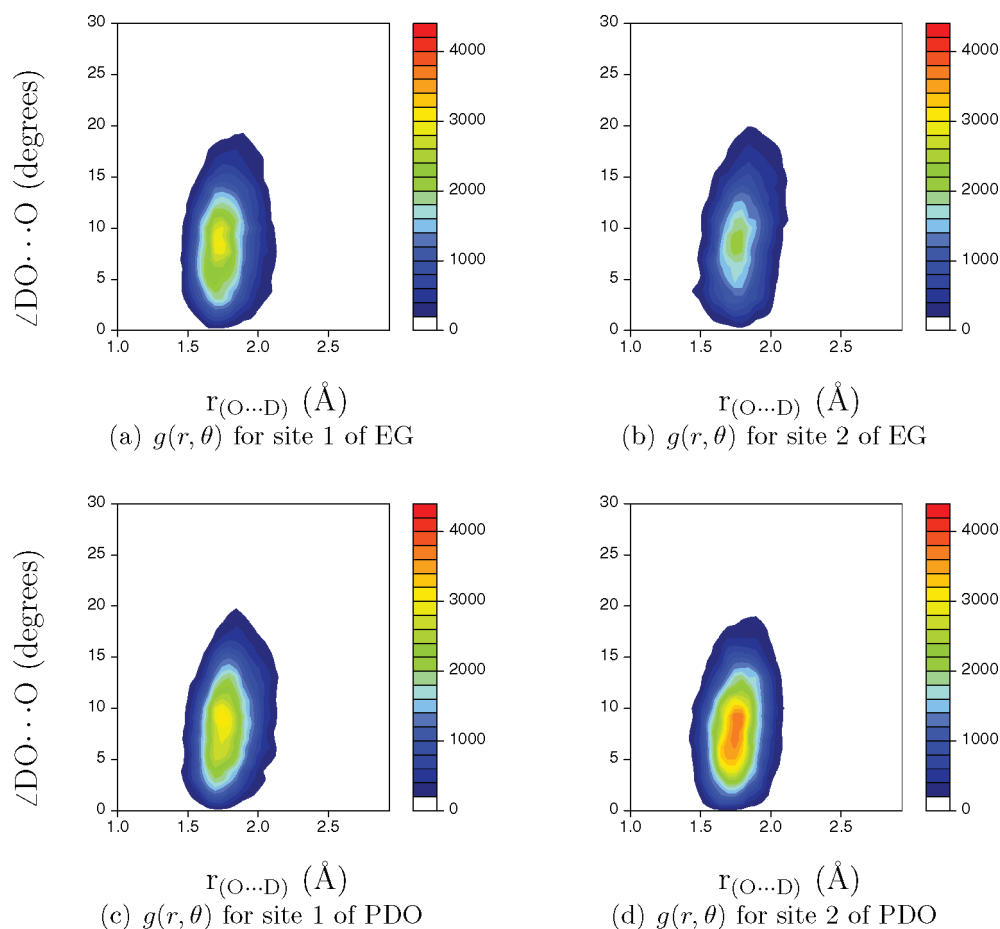


Figure 4. Weighted acceptor $g(r, \theta)$ for the two OD sites of the two diols.

while the corresponding effective frequency is proportional to $1/s$:

$$\nu = \frac{\omega_0 + \sqrt{2 + \omega_0^2}}{s4\pi} \quad (7)$$

For $\omega_0 = 2\pi$ (as used in this work), the previous expression reduces to $\nu \approx 1.01/s$. The parameter σ in eq 2 directly affects the time-frequency resolution, which is governed by a Heisenberg-like uncertainty principle.²⁹ In the present work, a value $\sigma = 2$ has been adopted in most spectrograms (unless otherwise specified) for a better time resolution. Similar choices for ω_0 (the main frequency) and σ parameters have been also adopted in refs 6 and 7. However, it must be pointed out that the choice of σ always represents a compromise between time and frequency resolution. This is shown in Figure 1 for a test function as

$$f_{\text{test}}(t) = \sin(kt + k') e^{-rt^2} \quad (8)$$

with k, k' , and r real constants, whose fast-Fourier transform (FFT) is the Gaussian enlarged “Dirac’s delta”-like function reported in blue in the inset of Figure 1. A roughly hyperbolic dependence of Γ (the width at half-height of the frequency peak of the test function) on σ is obtained, as shown in Figure 1.

When WT is used, it is possible to create a time-frequency plot. In the present work, where the main intermolecular interaction is the hydrogen bond between the solute diols and the solvent, we have analyzed both *intra*- and *intermolecular* structural properties. For the latter, defining the instantaneous bond length $r_{\text{O}\dots\text{D}}(t)$ between the O atom of a water molecule and the D

atom of the OD group of the solute, the time-resolved function $\Delta r_{\text{O}\dots\text{D}}(t) = r_{\text{O}\dots\text{D}}(t) - \langle r_{\text{O}\dots\text{D}}(t) \rangle$ has been taken as the fluctuation of the H-bond length. In the same manner, $\Delta r_{\text{O}-\text{D}}(t)$ is the fluctuation of the *intramolecular* O–D bond length of the diols. Equation 5, in these cases, produces a wavelet spectrogram that displays the change in time of the VDOS of the H-bond and O–D stretching mode, respectively.

The algorithm implemented in the present work looks for the value that maximizes the modulus $|\mathcal{W}'(n, s)|^2$ of the WT at a given time step n' . The corresponding value of $1/s$ is taken as the “instantaneous stretching frequency”. The same approach has been used to analyze *intramolecular* features by adopting as the input function the O–D bond length fluctuation of the hydroxyl groups of the glycols. The computer program developed in this work calculates the WT directly. A cutoff of $\pm 3\sigma$ from the maximum of the Gaussian Gabor-Morlet wavelet function has been used to allow the retention of more than 99.7% of the power spectrum energy and to save computation time.

3. RESULTS AND DISCUSSION

3.1. Structural and Electronic Properties. *3.1.1. Structural Properties.* PDO and EG interact with water mainly through H bonds, acting as both an acceptor and a donor. The hydrogen bond can be characterized structurally through the O–D \cdots O distance r and the O–D \cdots O angle θ . Figures 2 and 3 report the pair radial $g(r)$ and angular $g(\theta)$ distribution functions for the acceptor and donor H-bonds, respectively. It can be seen that

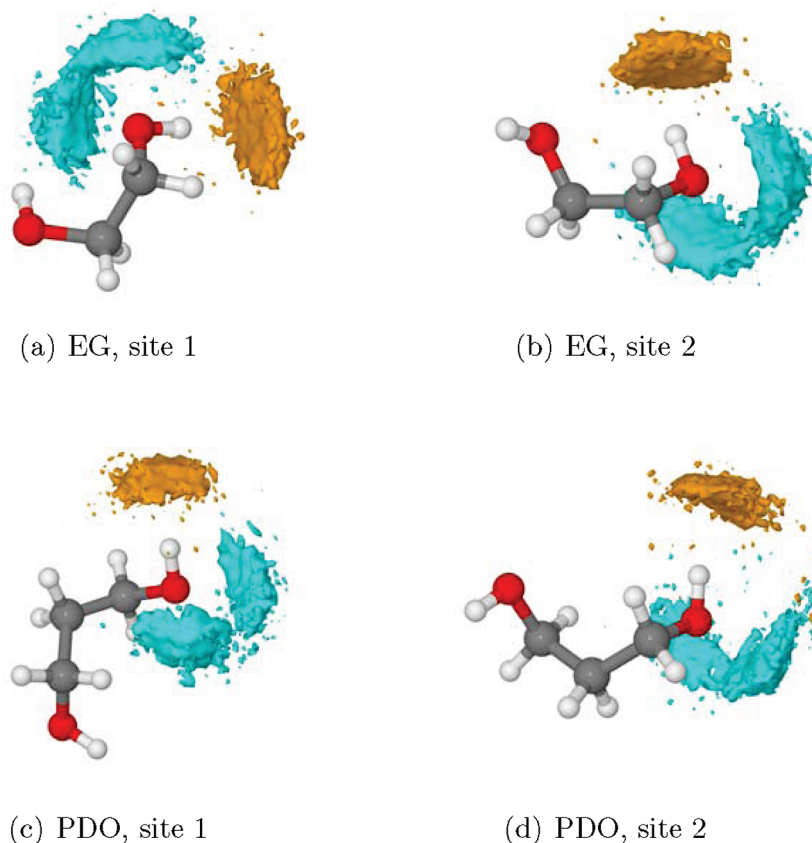


Figure 5. SDF of PDO and EG. Cyan and orange surfaces show the mobility of D and O atoms of H-bonded water molecules, respectively.

radial distribution functions of the two glycols differ very slightly. The deep first minima are indicative of the slow mobility of the water molecules bound to the glycols. On the contrary, it can be noted that the $g(\theta)$ functions for PDO are narrower, probably due to the fact that the two OD sites are farther away than in EG, allowing more independent solvation dynamics. This also implies that the angular parameters can be more sensitive to structural details. In the present case, the narrower $g(\theta)$ function is evidence for a more stable H-bond in the PDO solution.

The average coordination number for the OD sites, extracted from $g(r)$, is reported in Table 2 and is always close to the expected value of 1 for the donor interactions, whereas the acceptor coordination number is larger and shows a significant spread as a consequence of weaker interactions. The differences occurring for sites 1 and 2 of both diols arise from statistical uncertainty due to the finite temporal length of the simulations and the use of a single solute molecule for each system.

To better characterize the strength and structural features of the hydrogen bond, the weighted joint radial and angular distribution function $g(r, \theta)$ can be calculated by adopting the weighting function F_{HB} :^{30–32}

$$F_{\text{HB}} = A(r(t)) \times B(\theta(t)) \quad (9)$$

$$A(r(t)) = \begin{cases} \exp(- (r_e - r(t))^2 / 2\sigma_r^2) & \text{if } (r_e - r(t)) < 0 \\ 1 & \text{if } (r_e - r(t)) \geq 0 \end{cases} \quad (10)$$

Table 3. Average Dipole Moments for Diols According to MLWF Center Analysis on the Three Model Systems

	isolated diol	diol + H-bonded water molecules	diol in solution
EG	2.71 D	3.11 D	4.55 D
PDO	2.05 D	2.18 D	3.87 D

$$B(\theta(t)) = \begin{cases} \exp(- (\theta_e - \theta(t))^2 / 2\sigma_\theta^2) & \text{if } (q_e - q(t)) < 0 \\ 1 & \text{if } (q_e - q(t)) \geq 0 \end{cases} \quad (11)$$

$$g(r, \theta) = h(r) \times h(\theta) F_{\text{HB}} \quad (12)$$

The values of the r_e , θ_e , σ_r , and σ_θ parameters are extrapolated from unnormalized pair radial and angular distribution functions $h(r)$ and $h(\theta)$: r_e is the distance associated with the first local maximum in $h(r)$, and θ_e is the angle associated with the first local maximum of $h(\theta)$. σ_r and σ_θ are the half-widths at half-height in $h(r)$ and $h(\theta)$, respectively.

The results shown in Figure 4 confirm that H-bonding in PDO is stronger, due to the lower spread of the joint distribution function along the angular parameters, since in the present simulations the angular conditions are a more stringent requirement for the hydrogen bond formation. A pictorial view related to the weighted $g(r, \theta)$ is shown in Figure 5, where the spatial distribution functions (SDF) of the OD groups of the water molecules are reported, displaying a three-dimensional description

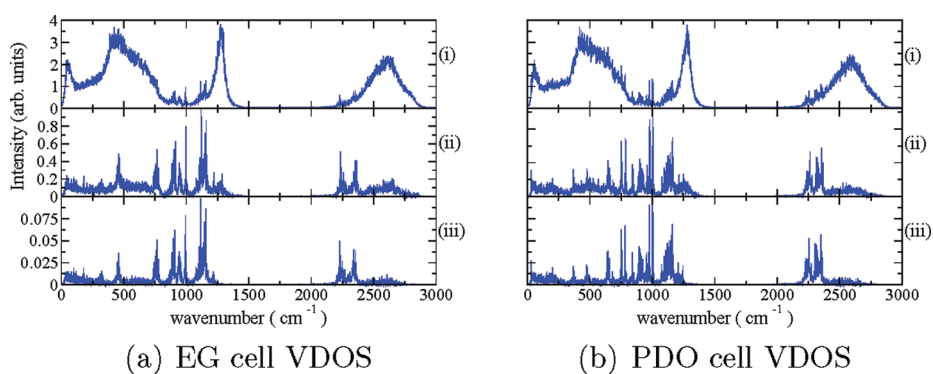


Figure 6. Vibrational densities of states: (i) VDOS of the complete simulation cells, (ii) VDOS of the diols and the two H-bonded water molecules, (iii) VDOS of the diols by themselves.

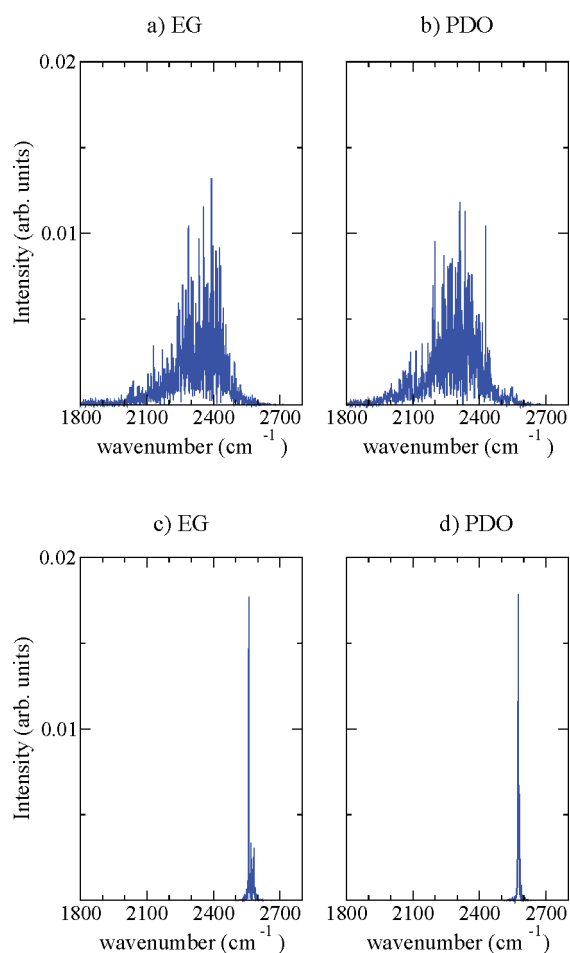


Figure 7. VDOS for hydroxyl group stretching mode in water (panels a and b) and in vacuum (panels c and d) simulations.

of the motion for the water molecules around the OD groups of the diols.

The cyan-colored isosurfaces (each point has been visited at least 55 times) close to the glycols represent the probability to find the D atoms of water around the O sites of the solutes, whereas the orange-colored ones represent the probability to find the O atoms of water around the alcoholic D atoms of the diols. These combined isosurfaces give a clear representation of the global mobility of the solvation cage. The spread of the

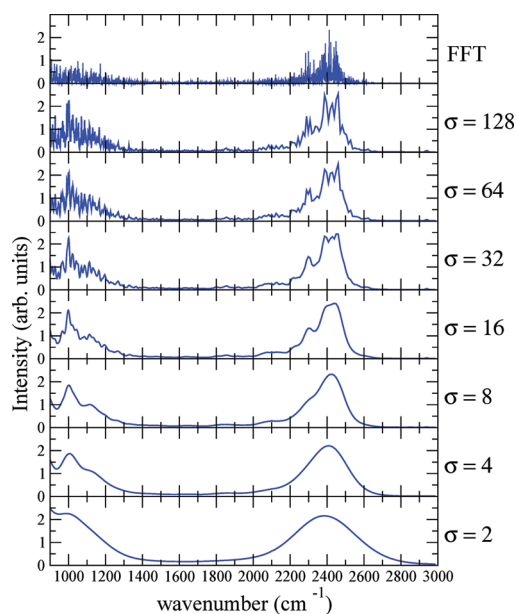


Figure 8. Fourier and wavelet power spectrum of the displacement of intermolecular $\Delta r_{O...D}(t)$ function of EG, site 2.

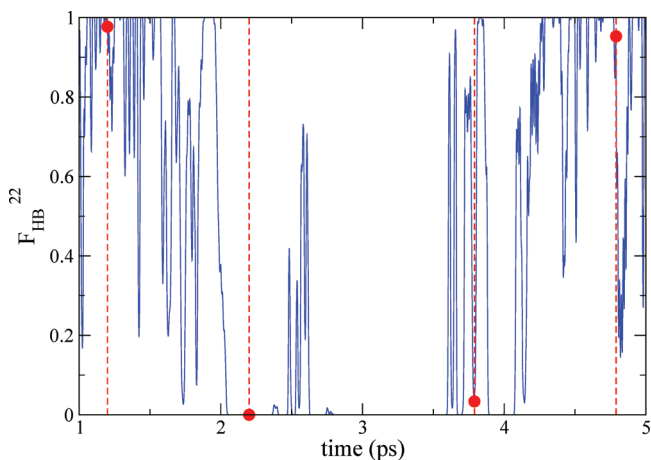


Figure 9. F_{HB}^{22} evolution with time for the interaction of D_2O molecule #22 with site 1 of PDO (full blue line). In dotted red lines are the sampled time steps for which the wavelet spectra are reported in Figure 10. The red circles represent the F_{HB}^{22} value sampled at those time steps.

isosurfaces for PDO is smaller than for EG, in line with the previous observations from the $g(\theta)$ and $g(r,\theta)$ functions.

3.1.2. Electronic Structure Analysis. The polarization effects due to the interaction of the glycols with the solvent have been analyzed in terms of Maximally Localized Wannier Functions (MLWF) centers.^{33,34} The positions of the MLWF centers can be related to the electron pairs and give a direct picture of the electronic structure. In order to obtain useful insight on the electronic structure changes due to the interactions of the glycols with the solvent, the molecular dipole moment computed on the basis

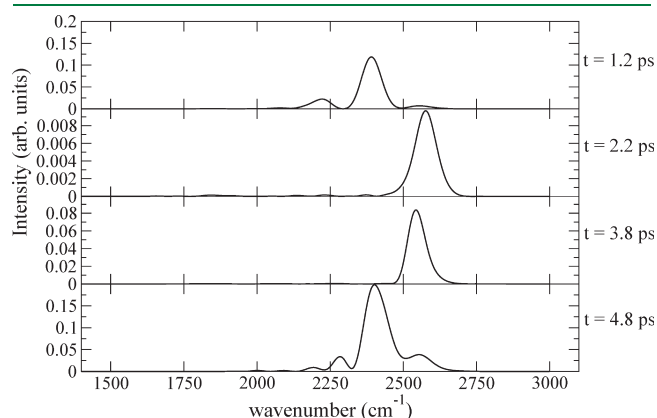


Figure 10. Wavelet spectra at four different time-steps calculated from the intramolecular $\Delta r_{\text{O-D}}(t)$ function for site 1 of PDO (parameter $\sigma = 10$).

of the MLWF centers positions of PDO and EG has been monitored during the simulations. For this purpose, the dipole moments of PDO and EG have been calculated on a series of time-equispaced configurations obtained by CPMD simulations (1 configuration every 10^4 time-steps, corresponding to 1.2 ps). The polarization enhancement of glycols in solution has been obtained by adopting the following computational strategy. For each configuration, the MLWF centers have been obtained for all of the system, for the diols and the H-bonded water molecules and for diols alone without geometry optimization.

In Table 3, the dipole moments calculated by MLWF center analysis by localizing the valence shell electron doublets are reported. The dipole moment of the diols increases as reported in Table 3, as a consequence of the interaction with the solvent. The time dependent data also show an evident correlation in the time evolution of the dipole moments for the three model systems. However, the solvent increases the dipole moment of the diols, and larger fluctuations are observed for EG, which being smaller in size is more sensitive to polarization effects by the water cage. Furthermore, the dipole moment increase, due to solvent interaction, is larger in EG than in PDO for the same reason.

3.2. Time-Frequency Analysis. In Figure 6, the VDOS of the water solutions, of the diols with the two closest molecules, and of the diols by themselves in the water solutions are shown, obtained by FFT of the velocity autocorrelation functions. The OD stretching region ($2100\text{--}2600\text{ cm}^{-1}$), which is the most interesting region of the spectrum, will be discussed in the following on the basis of WT analysis.

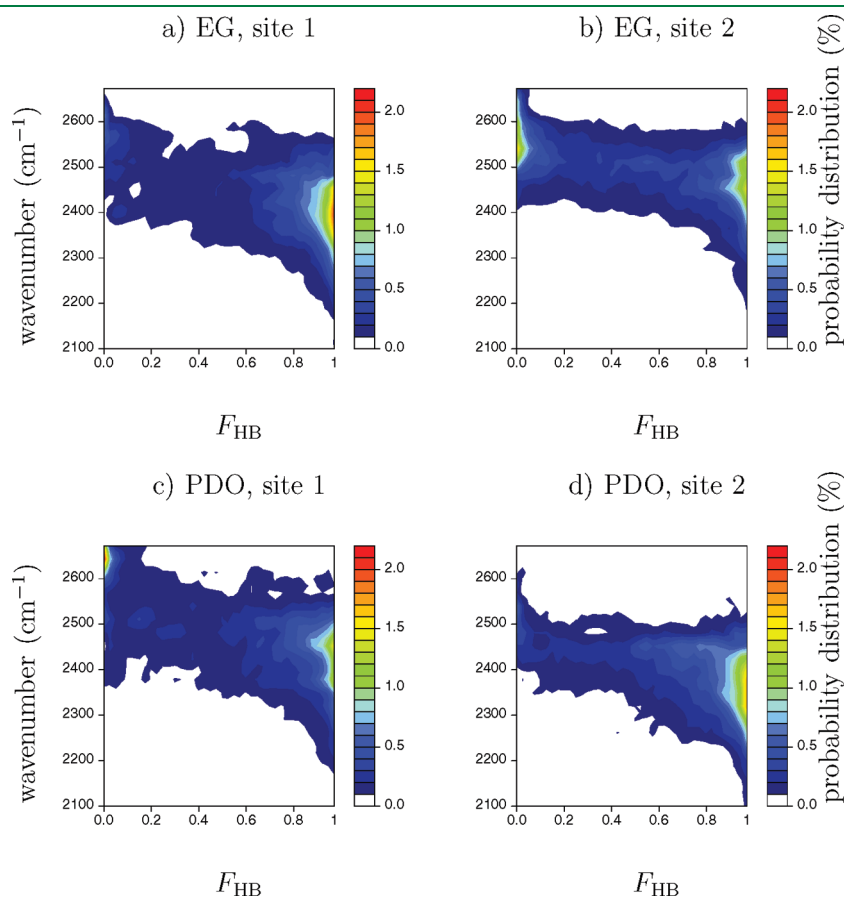


Figure 11. Correlation spectrograms between O–D stretching frequency and F_{HB} function.

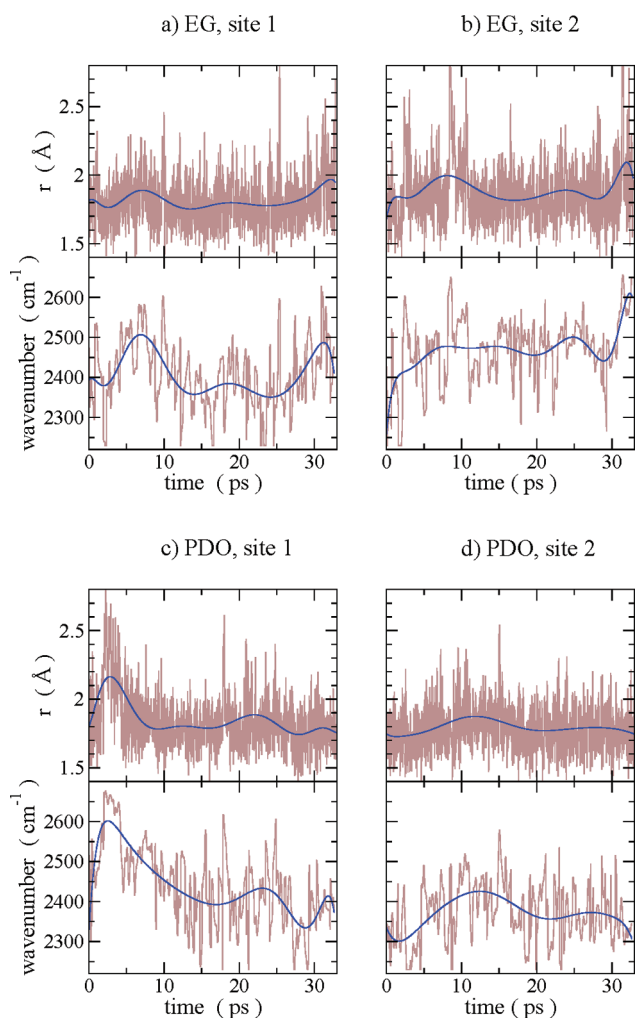


Figure 12. Comparison of the time evolutions of H-bond length and O–D stretching frequency. The brown lines correspond to the actual values found during the simulations, whereas the blue lines represent the smoothed trends of the former.

The structure of the OD stretching band results from the convolution of solvent–solvent and solute–solvent interactions, and we focus on the structure, extracted from the simulation, of the VDOS of the solute molecule and of the two nearest water molecules bound to the diols. This can be appreciated in more detail when analyzing the VDOS extracted by FFT of the oscillations of the intramolecular $\Delta r_{\text{O–D}}$ functions, as reported in Figure 7, where the density of states obtained by simulations of the diols in solution (panel *a* and *b*) and in a vacuum (panel *c* and *d*) are reported. The red shift due to H-bonding interactions with the water solvent is evident, along with the broadening of the stretching band.

In Figure 8, the FFT of this feature of the simulation cell is compared with the WT at various values of the σ parameter showing details of the band structure at increasing resolution, using the intermolecular $\Delta r_{\text{O}\cdots\text{D}}(t)$ fluctuation as an input function.

The WT appears to correctly reconstruct the Fourier spectrum with high values of parameter σ , showing that the truncation in the Gabor-Morlet wavelet does not significantly affect the calculations. The inhomogeneous broadening of the band can be attributed to time changes of the structure of the solvent cage around the hydroxyl groups of the diols, which modulates the instantaneous vibrational frequency.

To better enlighten the origin of the various frequency components, the trajectory of site 1 of PDO has been analyzed by WT at specific time intervals. Figure 9 shows the details of the trajectory referring to site 1 of PDO. It can be seen that along the first 5 ps of simulation, site 1 is alternatively free or bound to water molecule #22. Later, along the trajectory, the site is constantly bound to the same water molecule.

The wavelet transform spectrum has been calculated at the simulation times of 1.2, 2.2, 3.8, and 4.8 ps (dotted red lines in Figure 9), and the results, reported in Figure 10, clearly show that the red shift nicely correlates with the H-bonding character. The same correlation has been observed at all time-steps probed for both sites of the two glycols.

An alternative and more immediate way of illustrating this behavior is obtained by directly correlating the most intense O–D stretching frequency, obtained by WT of the corresponding $\Delta r_{\text{O–D}}(t)$ function, with the values of the F_{HB} donor function. This is displayed in Figure 11 showing how the wavelet-calculated VDOS for all OD sites of the diols changes with the value of the F_{HB} function (which is confined in the $[0-1]$ interval due to the fact that only a single donor H-bond can be present at a time). The plotted quantities are the probability distributions of the VDOS, which thicken and are red-shifted when F_{HB} approaches 1, meaning that the intermolecular H-bond lowers the O–D intramolecular stretching frequency. Consistently, the VDOS is peaked at higher frequencies when $F_{\text{HB}} \approx 0$. As can be seen, there is a sort of pathway between the two extremes in the frequency/configurational space explored by the simulations. For site 2 of both glycols, the pathway is continuous, due to the more oscillating character of the F_{HB} function with oscillations too fast to be precisely resolved even with WT.

As a whole, the present analysis shows how a structural quantity like F_{HB} , designed as a probe of the H-bonding, can be efficiently correlated with the stretching frequency of an intramolecular vibration, because the molecular oscillators are coupled to the water environment and are sensitive to small fluctuations in the solvation cage. Moreover, the advantage of a continuous function like F_{HB} to monitor on the fly the dynamics of the hydrogen bond and its effect on the vibrational spectrum is quite evident.

Wavelet analysis has also been used to sort out the time evolution of the frequency related to the most intense peak in the high resolution spectra (obtained like those of Figure 10) and its correlation with the intermolecular O \cdots D bond length between nearest molecules along the simulation. The result is shown in Figure 12. The blue line in Figure 12, drawn as a guide for the eye, has been obtained by a 10th-order polynomial fit of the raw data.

A straight correlation between the time evolution of vibrational frequencies and the time evolution of intramolecular O–D stretching distance can be appreciated.

A more straightforward way to show the correlation between stretching frequencies and intermolecular $\text{D}_{\text{diol}}\cdots\text{O}_{\text{water}}$ distance is via the bidimensional spectrogram plot of Figure 13, where distance and frequency are directly correlated at each time step. The probability distribution reported in Figure 13 represents the maximum of the intramolecular O–D band stretching obtained by WT correlated with the corresponding value of the intermolecular H-bond length.

These spectrograms show the same “banana shaped” distributions (as referred to in refs 6 and 7) obtained for other systems like heavy water and Cl^- anions in heavy water. Moreover, the same type of shape is obtained using intermolecular H-bond distances or intramolecular O–D bond lengths, since they both

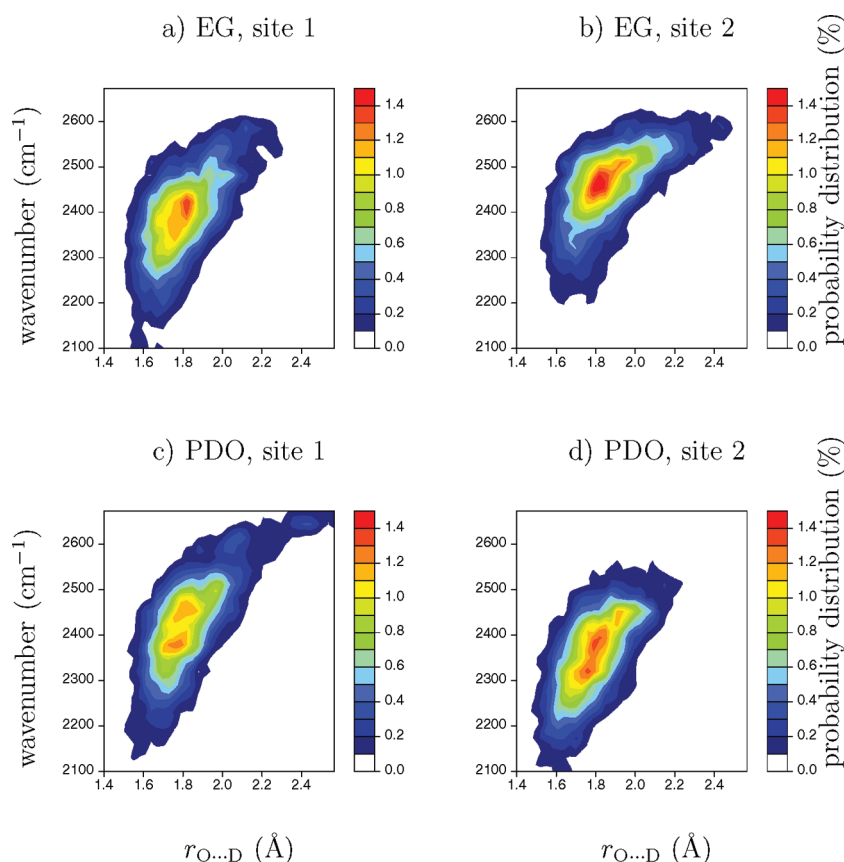


Figure 13. H-bond length–frequency spectrograms for both PDO and EG.

mainly probe the stretching frequency associated with the O–D stretching of the glycols. The VDOS thickens at bond lengths corresponding to the first peak of the $g(r)$ and $g(r, \theta)$ functions. The frequency shift is thus simply correlated to the strength and stability of the H-bond. A strong H-bond gives a $g(r)$ function peaked at shorter distances, and within the harmonic approximation, it reduces the intramolecular stretching force constant and, consequently, shifts the VDOS at smaller wavenumbers. The slightly different plot for EG, site 2, is due to the fact that the most strongly interacting water molecule for that site is not H-bonded for the first picosecond of simulation, thus resulting in a final spectrogram showing a weaker H-bond, as expected.

It is not possible to establish an exact bijective relationship between bond length and frequency, but this is not surprising because the radial distance is only one of the structural parameters that characterize the H-bond, the other being the angular parameter. Hence, the distance is degenerate in the angle space. Moreover, the Heisenberg-like uncertainty principle between time and frequency, and therefore between distance and frequency, affects the spectrogram resolution. However, these plots clearly show that a reliable correlation between bond length and density of vibrational states exists, which can be understood on the basis of the H-bond stability. This is in agreement with previous results obtained using static *ab initio* calculations.³⁵

4. CONCLUSIONS

In the present work, the hydrogen bond structure and dynamics of two prototype glycols, ethylene glycol and propanediol, in water solutions have been studied by *ab initio* molecular

dynamics simulations in the density functional theory approach. Molecular dynamics trajectories have been obtained using the Car–Parrinello (CPMD) method and have been analyzed by both the traditional Fourier transform and the wavelet transform methods. It has been shown that a good deal of novel information can be obtained by the time analysis supplied by the wavelet analysis. In particular, it has been found that the complex pattern of the hydrogen bond stretching mode has an inhomogeneous origin and arises from the convolution of a number of differently shifted vibrational modes that can be sorted out by wavelet transform at different time intervals of the molecular dynamics simulation. The time-resolved vibrational modes obtained by the wavelet analysis can be correlated in a straightforward way with the structure of the solvent cage around the solute diols and with structural parameters like the O–D bond length or with the strength of the hydrogen bonding. A careful comparison of the Fourier and wavelet transforms also shows that for a more accurate characterization of the hydrogen bonding a simultaneous consideration of the bond length (e.g., $r_{\text{O–D}\cdots\text{O}}$) and of the $\angle\text{O–D}\cdots\text{O}$ angle is needed. The wavelet transform analysis is not only a suitable tool to analyze the vibrational features in terms of modulation of the local structure of the solvation cage but also offers unique opportunities to study time-resolved spectroscopic experiments.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: gianni.cardini@unifi.it.

ACKNOWLEDGMENT

This work was supported by the Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR). The authors would like to thank the CINECA Supercomputing Center for the allocation of computing resources and Prof. Roberto Righini (LENS) for useful discussions.

REFERENCES

- (1) Goedecker, S.; Ivanov, O. V. *Phys. Rev. B* **1999**, *59*, 7270–7273.
- (2) Neelov, A. I.; Goedecker, S. *J. Comput. Phys.* **2006**, *217*, 312–339.
- (3) Daykov, I. P.; Arias, T. A.; Engeness, T. D. *Phys. Rev. Lett.* **2003**, *90*, 216402–216405.
- (4) Engeness, T. D.; Arias, T. A. *Phys. Rev. B* **2002**, *65*, 165106–165115.
- (5) Cho, K.; Arias, T. A.; Joannopoulos, J. D.; Lam, P. K. *Phys. Rev. Lett.* **1993**, *71*, 1808–1811.
- (6) Mallik, B. S.; Semparathi, A.; Chandra, A. *J. Chem. Phys.* **2008**, *129*, 194512–194527.
- (7) Mallik, B. S.; Semparathi, A.; Chandra, A. *J. Phys. Chem. A* **2008**, *112*, 5104–5112.
- (8) Askar, A.; Cetin, A. E.; Rabitz, H. *J. Phys. Chem.* **1996**, *100*, 19165–19173.
- (9) Rahaman, A.; Wheeler, R. A. *J. Chem. Theory Comput.* **2005**, *1*, 769–771.
- (10) Pagliai, M.; Muniz-Miranda, F.; Cardini, G.; Righini, R.; Schettino, V. *J. Phys. Chem. Lett.* **2010**, *1*, 2951–2955.
- (11) Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471–2474.
- (12) Orts, W. J.; Holtman, K. M.; Seiber, J. N. *J. Agric. Food Chem.* **2008**, *56*, 3892–3899.
- (13) CPMD; MPI für Festkörperforschung Stuttgart: Stuttgart, Germany, 1997–2001; IBM Corp.: Armonk, New York, 1990–2008.
- (14) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.
- (15) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.
- (16) Troullier, N.; Martins, J. L. *Phys. Rev. B* **1991**, *43*, 1993–2006.
- (17) Kleinman, L.; Bylander, D. M. *Phys. Rev. Lett.* **1982**, *48*, 1425–1428.
- (18) Kuo, I. F. W.; Mundy, C. J.; McGrath, M. J.; Siepmann, J. I.; VandeVondele, J.; Sprik, M.; Hutter, J.; Chen, B.; Klein, M. L.; Mohamed, F.; Krack, M.; Parrinello, M. *J. Phys. Chem. B* **2004**, *108*, 12990–12998.
- (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. J.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347–1363.
- (20) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (21) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372–1377.
- (22) Howard, D.; Jorgensen, P.; Kjaergaard, H. G. *J. Am. Chem. Soc.* **2005**, *127*, 17096–17103.
- (23) Kinning, A.; Mom, V.; Mijlhoff, F. C.; Renes, G. H. *J. Mol. Struct.* **1982**, *82*, 271–275.
- (24) Shao, X. G.; Leung, A. K. M.; Chau, F. T. *Acc. Chem. Res.* **2003**, *36*, 276–283.
- (25) Ehrentreich, F. *Anal. Bioanal. Chem.* **2002**, *372*, 115–121.
- (26) Carmona, R.; Hwang, W.; Torresani, B. *Practical Time-Frequency Analysis: Gabor and Wavelet Transforms with an implementation*; Academic Press: New York, 1998.
- (27) Kirby, J. F. *Comput. Geosci.* **2005**, *31*, 846–864.
- (28) Torrence, C.; Compo, G. P. *Bull. Am. Meteorol. Soc.* **1998**, *79*, 61–78.
- (29) Chui, C. K. *An introduction to wavelet; Wavelet analysis and its applications*; Academic Press, Inc.: New York, 1992.
- (30) Pagliai, M.; Cardini, G.; Righini, R.; Schettino, V. *J. Chem. Phys.* **2003**, *119*, 6655–6662.
- (31) Pagliai, M.; Cardini, G.; Schettino, V. *J. Phys. Chem. B* **2005**, *109*, 14923–14928.
- (32) Faralli, C.; Pagliai, M.; Cardini, G.; Schettino, V. *J. Phys. Chem. B* **2006**, *110*, 14923–14928.
- (33) Marzari, N.; Vanderbilt, D. *Phys. Rev. B* **1997**, *56*, 12847–12862.
- (34) Silvestrelli, P. L.; Marzari, N.; Vanderbilt, D.; Parrinello, M. *Solid State Commun.* **1998**, *107*, 7–11.
- (35) Klein, R. A. *J. Comput. Chem.* **2002**, *23*, 585–599.

A Stochastic Search for the Structures of Small Germanium Clusters and Their Anions: Enhanced Stability by Spherical Aromaticity of the Ge_{10} and Ge_{12}^{2-} Systems

Truong Ba Tai[†] and Minh Tho Nguyen^{*,†,‡}[†]Department of Chemistry, and LMCC-Mathematical Modeling and Computational Science Center, Katholieke Universiteit Leuven, B-3001 Leuven, Belgium[‡]Institute for Computational Science and Technology, Thu Duc, Ho Chi Minh City, Vietnam Supporting Information

ABSTRACT: Investigations on germanium clusters in the neutral, anionic, and dianion states Ge_n^x ($n = 2-12$ and $x = 0, -1, -2$) are performed using quantum chemical calculations with the B3LYP functional and the coupled-cluster singles and doubles [CCSD(T)] methods, in conjunction with the 6-311+G(d) basis set. An improved stochastic method is implemented for searching the low-lying isomers of clusters. Comparison of our results with previous reports on germanium clusters shows the efficiency of the search method. The Ge_8 system is presented in detail. The anionic clusters $\text{Ge}_n^{-/2-}$ are studied theoretically and systematically for the first time, and their energetics are in good agreement with available experiments. The clusters Ge_{10} , Ge_{10}^{2-} , and Ge_{12}^{2-} are, in their ground state, characterized by large highest occupied molecular orbital–lowest unoccupied molecular orbital gaps, high vertical and adiabatic detachment energies, and substantial average binding energies. The enhanced stability of these magic clusters can consistently be rationalized using the jellium electron shell model and the spherical aromatic character.

1. INTRODUCTION

Germanium (Ge) clusters have attracted much interest in part due to their possible role in surface growth processes and applications in electronic industries as alternatives to silicon-based materials. Ge thin films are in fact considered for semiconductors, and the deposition of Ge layers is generally achieved by chemical vapor deposition using germane.¹ The first experimental detection of Ge clusters containing 2–8 atoms dates back to 1954.² Subsequently, numerous experimental and theoretical investigations on the small-to-medium sized Ge_n were reported. We would refer to the most recent reports^{3–10} for the numerous earlier references. The pure Ge clusters are known to be chemically reactive and, therefore, not quite suitable as building blocks of self-assembly materials.¹¹ By using an appropriate dopant, it is however possible to modify the cluster chemical properties and thereby to design new and relevant materials. Recently, we have investigated the Cr- and Li-doped germanium clusters (Ge_nCr and Ge_nLi_m)^{12–16} and demonstrated both experimentally and theoretically that the Li-doped derivatives can actually form nanowires from units of Ge_9Li_3 .¹⁶ In small systems, the Li elements are found to attach exclusively outside the Ge cores and undergo electron transfer, and as a consequence, the Ge–Li bonds are essentially of ionic character. In Ge_{12}Li , Li can be placed inside the Ge cage, and the endohedrally doped cluster turns out to be the most stable form.¹⁷ In this context, the Ge cores of the smaller doped clusters can best be regarded as anions or polyanions Ge_n^{x-} , with $n < 12$, interacting by electrostatic forces with Li cations. Despite the fact that experimental detections of anionic clusters Ge_n^- have long been reported,^{18,19} relatively less is known about their structures and stabilities as compared with their neutral counterparts. To the

best of our knowledge, only a few theoretical studies have been devoted to the anionic clusters Ge_n^- , with n up to 8. Deutsch et al.²⁰ carried out G2 and density functional theory (DFT) calculations for the electron affinities and binding energies of Ge_n^- ($n = 2-5$). Studies of the monoanions Ge_n^- ($n = 2-6$) were performed by Xu et al.²¹ and Archibong et al.²² using the B3LYP and coupled-cluster singles and doubles [CCSD(T)] methods. In a series of theoretical investigations on the germanium clusters, both neutral and ionic Ge_n^x ($n = 5-8$ and 12 and $x = +4, +2, 0, -2, -4$), King and co-workers²³ recently proposed a geometrical analogy between the dianions Ge_n^{2-} , with $n = 5-7$, and boron hydrides $\text{B}_n\text{H}_n^{2-}$. However, this tendency does not hold true for Ge_8^{2-} where a tetrahedral structure Ge_8^{2-} (T_d) is the global minimum instead of the D_{2d} structure of $\text{B}_8\text{H}_8^{2-}$.²⁴ A similar failure was also found for system containing 12 atoms in that a C_{2v} structure is the lowest-lying Ge_{12}^{2-} isomer,²⁵ whereas an icosahedron (I_h) is well established as the global minimum of $\text{B}_{12}\text{H}_{12}^{2-}$.²⁶

In view of the lack of reliable information on the anionic Ge clusters, we set out to perform a systematic study of their anionic and dianionic states. In order to evaluate the stabilities of the anions with respect to electron detachment, the corresponding neutral systems are also considered. In this study, the lower-lying isomers of clusters are initially searched for using a modified stochastic method, implemented by us,²⁷ in conjunction with reliable electronic structure methods.

Received: November 10, 2010

Published: March 07, 2011

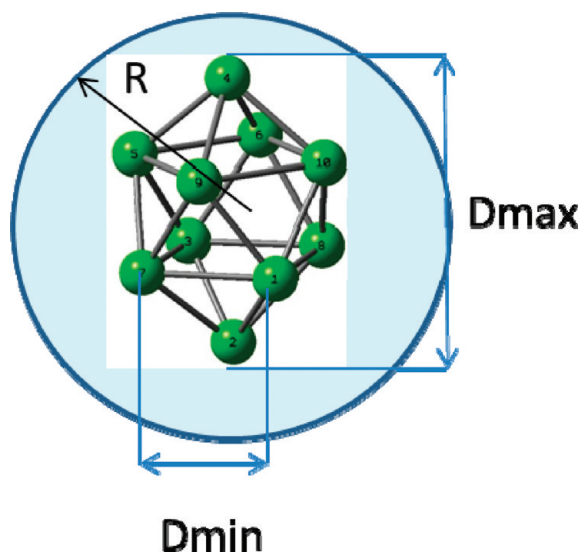


Figure 1. Control variables of the kick procedure.

2. COMPUTATIONAL METHODS

Stochastic Search for Structures. The search for the energetically lower-lying geometrical structures for large chemical compounds remains a challenge for theoretical chemists. Many search methods using various computational techniques have been proposed. For instance, the genetic algorithm-based methods have been used for searching the minima of clusters by different research groups.^{28,29} A dynamic simulated annealing method that combines molecular dynamics and DFT has widely been employed.³⁰ A ‘minimum hopping’ method used to scan the entire energy potential surface has also been put forward.³¹

Saunders³² reported a ‘random kick’ procedure for searching the low-lying isomers of compounds, which is a derivative of the stochastic search method. In this random kick procedure, each atom of an initial structure is kicked to randomly move within a sphere of radius r , and the structures so constructed will be the inputs for following geometry optimizations using electronic structure calculations. The only variable controlled here is the ‘moving radius’ r of atoms. Although this procedure was proven to be effective for searching the global minima of compounds in recent reports, the simplicity of the procedure tends to lead to a small ratio of structures effectively converged after optimization. The structures obtained by kick in which the distance between atoms is either too short or too long are either not converged or fragmented into many smaller species upon optimization. The yield of completed optimization ranges only from 10 to 50% of the total structures generated.³³

In this work, we implement a modified stochastic searching procedure that allows the yield of completed optimization up to 90% to be attained. Similar to the original Saunders’ procedure, each of the atoms of an initial structure is kicked to randomly move within a kick radius r . However, three additional variables will be controlled to provide better structures constructed for the subsequent geometry optimization. The first variable is the distance from kicked atoms to the center of structure, called the maximum radius of molecule R (Figure 1). The next variables are the minimum distance D_{\min} , and the maximum distance D_{\max} between atoms. It is noted that, while the first variable is used to limit the movement of atoms and to make the kick procedure

performed more easily, the latter variables play an important role in the structure construction.

Our search for the lower-lying isomers of Ge_n is thus carried out using the improved stochastic method outlined above in conjunction with geometry optimizations by electronic structure calculations using the Gaussian 03 program package.³⁴ An input file containing the coordinate of any initial structure of Ge_n is subject for kick procedure is given in the Supporting Information. The i^{th} atom of a structure with initial coordinates (x_i, y_i, z_i) is kicked to randomly move within a sphere of radius r to new position $(x_i + dx, y_i + dy, z_i + dz)$. At this stage, three variables are controlled to obtain a new structure for a geometry optimization. First, the distance from the ‘kicked’ atom to the center of structure $[\{(x_i + dx)^2 + (y_i + dy)^2 + (z_i + dz)^2\}]^{1/2}$ must be smaller than the maximum radius R of molecule. Second, the distance between any pair of i^{th} and j^{th} atoms as defined by $[\{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2\}]^{1/2}$ must be in the range of $D_{\min} - D_{\max}$. If both conditions are satisfied, then a new structure will then be obtained and is considered for geometry optimization using an electronic structure method. Otherwise, the structure will be rejected, and a new ‘kick’ is carried out. In general, the coordinates of the i^{th} structure obtained by the kick procedure will be used as initial coordinates for the $(i + 1)^{\text{th}}$ kick. A chart for the whole procedure is given in Scheme S1 of the Supporting Information.

A random structure for each cluster size is used as an initial coordinate for the kick procedure. Maximum radius R of molecules is chosen from 5 to 7 Å, being approximately 2–3 times as large as the bond length of the diatomic Ge_2 . The minimum distance D_{\min} between two atoms is set to be 2 Å, which is slightly smaller than bond length of Ge_2 , while the maximum distance D_{\max} varies within the range of 5–8 Å.

Electronic Structure Methods. The structures generated for each size are initially optimized using the popular B3LYP hybrid functional of DFT,^{35,36} along with the small 6-31G basis set. The lower-lying minima, obtained from these optimizations, characterized by harmonic vibrational frequencies at the same level, are further refined using a higher level of theory. The geometry optimization and calculation of harmonic vibrational frequencies for the located local minimum structures are further performed using the B3LYP/6-311+G(d)³⁷ level. In order to evaluate the accuracy of the calculated DFT results, single point electronic energy calculations of the smaller clusters Ge_m , with $n = 2-6$, are carried out using the CCSD(T) method,³⁸ in conjunction with the correlation consistent aug-cc-pVTZ basis set.³⁹ The molecular orbital (MO) analyses of the global minima systems are performed using the B3LYP/6-311+G(d) densities. Comparisons are made with, where appropriate, available experimental values. In addition, the stability features of clusters are considered in terms of the jellium shell model (JSM),⁴⁰ which was applied successfully on metal clusters.^{17,41}

3. RESULTS AND DISCUSSION

Efficiency of Search Method. Up to 100 structures for each size of Ge_n , in particular for $n = 6-12$, are selected using the kick procedure and initially optimized at the B3LYP/6-31G level. Due to limitation of computational resources, only the lower-lying isomers having relative energies smaller than 50 kcal/mol with respect to the global minimum are further refined using the B3LYP/6-311+G(d) level. The shapes of the lowest-lying isomers are shown in Figures 3 and 4.

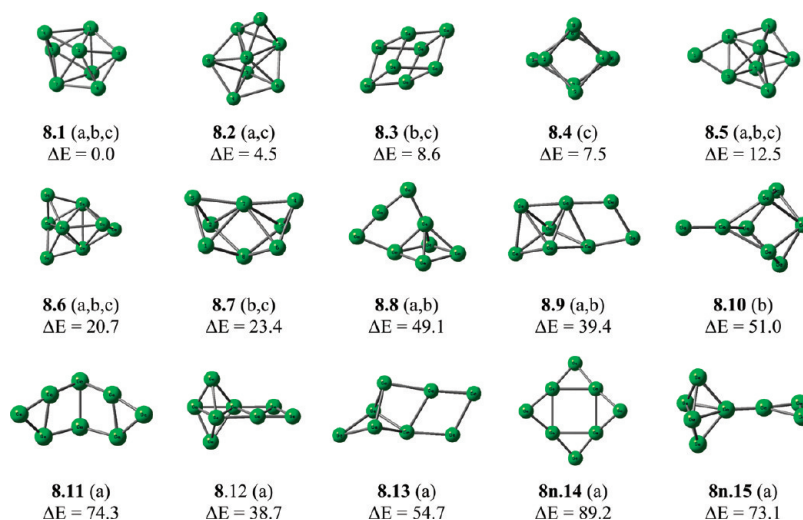


Figure 2. Shapes and relative energies (kcal/mol) of the Ge_8 isomers obtained using three different kick procedures. (a) Isomers obtained using the kick procedure without limitation of distance D ; (b) Isomers obtained using the kick procedure with $D = 2\text{--}9 \text{ \AA}$; and (c) Isomers obtained using the kick procedure with $D = 2\text{--}5 \text{ \AA}$.

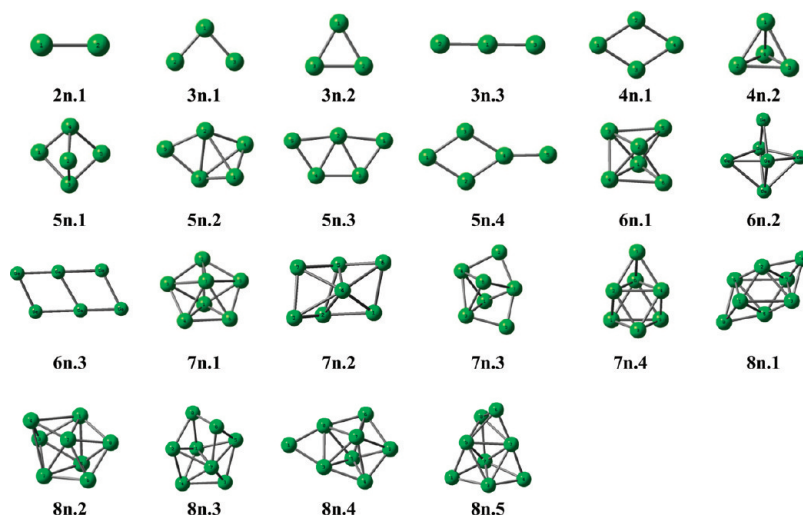


Figure 3. Shapes of the low-lying isomers Ge_n ($n = 2\text{--}8$).

Let us note that the search for structures of the small clusters Ge_n ($n \leq 7$) only results in a few novel isomers because the number of isomers of this series is rather limited, and they have already been reported. For example, for Ge_6 a new triangular prism structure is located in our search, but frequency calculations show that this structure is only a second-order saddle point having two imaginary frequencies. Therefore it is not given here. Thus, in order to evaluate the efficiency of searching method and the effect of the controlled distance D between atoms on the efficiency of the kick method, the search for Ge_8 structures is considered in more detail using three different kick procedures. All Ge_8 isomers located are given in Figure 2. In the first procedure, the kick procedure is performed without a limitation on D . With 100 structures thus obtained, the yield of completed optimization is 46%. This number rises up to 92 and 76% when Ge_8 structures are generated by kick procedures in which the distance D is limited from $D_{\min} = 2 \text{ \AA}$ to $D_{\max} = 5 \text{ \AA}$ and to $D_{\max} = 9 \text{ \AA}$, respectively. Additionally, it can be seen that when distance D is not limited, the potential energy surface scanned is larger. Thus, some less stable isomers can be located, as shown in

Figure 2, whereas some stable isomers are missed. The stable isomers with small relative energies are obtained when this distance is controlled at suitable values. Consequently, a limitation of distance D between atoms thus makes the optimization yield higher, and the structure search becomes more effective.

Similarly, a search for the low-lying isomers of the larger clusters Ge_n ($n = 9\text{--}12$) is performed, and the results are summarized Figure 4. The optimization yield varies from 80 to 90%. The local minima obtained are overall in agreement with previous reports.^{4,7}

Low-Lying Isomers of Ge_n Clusters. Characterization of the structures and stabilities of the neutral clusters Ge_n has been made. For instance, Ho and co-workers⁷ performed a search for the global minima for small clusters Ge_n ($n \leq 16$) using the combination of a genetic algorithm with an empirical tight-binding method. Wang et al.⁴ reported the low-lying isomers of medium-sized clusters ($n = 2\text{--}25$) determined by combining a genetic algorithm with a nonorthogonal tight-binding method. More recently, Zeng and co-workers⁹ searched for global minima of small-to-medium-sized germanium clusters Ge_n ($n = 12\text{--}29$) by

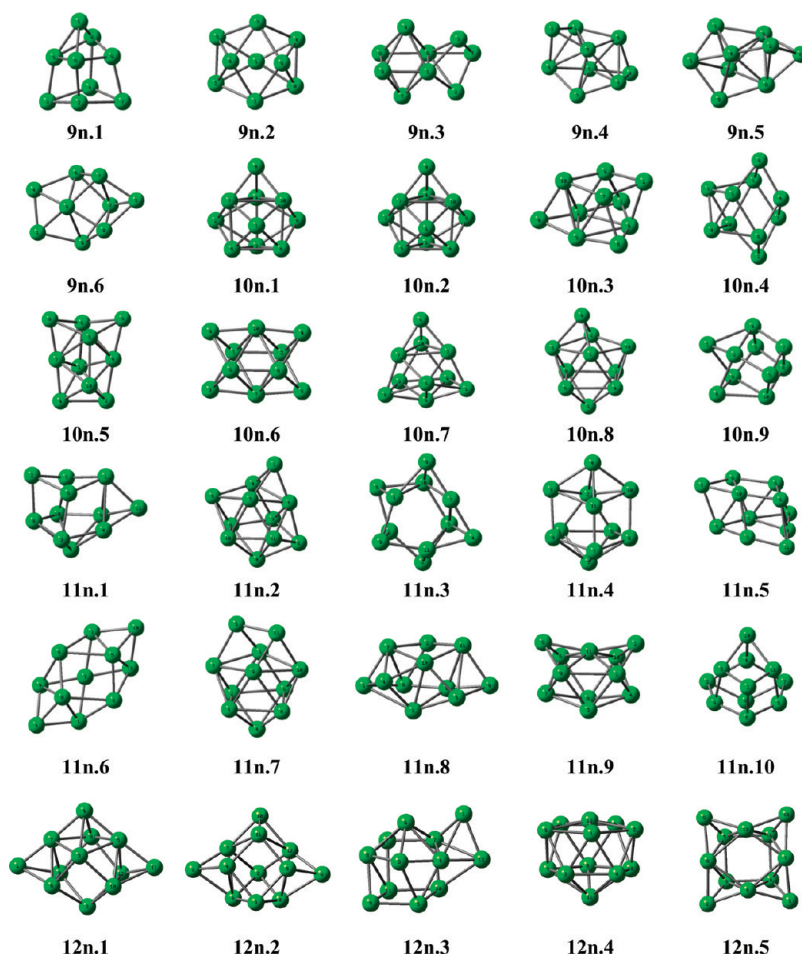


Figure 4. Shapes of the low-lying isomers Ge_n ($n = 9-12$).

using a basin-hopping algorithm coupled with the plane-wave pseudopotential density functional calculations. Although these theoretical studies have not always agreed with each other on the identity of the most stable forms in some large systems, the global minima of the small Ge clusters are now well established.

While the shapes of the low-lying isomers Ge_n^x at neutral, anionic, and dianionic states are shown in Figure 3–6, their point group, electronic state, and relative energies are given in Tables 1 and 2. In the Figures as well as in the following Discussion Section, the relative energies given are evaluated with respect to the global minimum in each isomeric system. In good agreement with earlier results,^{4,6–8} the structures **2n.1** ($^3\Sigma_g^-$), **3n.1** (C_{2v} , 1A_1), **4n.1** (D_{2h} , $^1A_1'$), **5n.2** (D_{3h} , $^1A_1'$), **6n.1** (C_{2v} , 1A_1), and **7n.1** (D_{5h} , $^1A_1'$) (Figure 3) are found to be the global minima of Ge_2 , Ge_3 , Ge_4 , Ge_5 , Ge_6 , and Ge_7 , respectively. However, some high-stability isomers are also found in our search. Two structures **3n.1** and **3n.2** are almost degenerate with a negligible energy difference of 0.5 kcal/mol at the B3LYP/6-311+G(d) level. This gap slightly increases to 1.4 kcal/mol at the CCSD(T)/aug-cc-pVTZ level. Thus we prefer to assign **3n.1** as a global minimum for Ge_3 . Similarly, two structures **6n.1** and **6n.2** are found to be degenerate with relative energy of only 0.3 kcal/mol at the CCSD(T) level. At the B3LYP/6-311+G(d) + zero point energy (ZPE) level, we found that two structures **8n.1** (C_{2v} , 1A_g) and **8n.2** (C_s , $^1A'$) are almost degenerate. Our CCSD(T) results show a similar ordering in which **8n.1** is 1.2 kcal/mol more stable than **8n.2**. The next isomer is **8n.3** (C_{2v} , 1A), being only 3.0 kcal/mol higher in energy. Both isomers **8n.4** (C_s , $^1A'$) and **8n.5**

(C_1 , 1A) also turn out to be stable, as they are 4.8 and 10.9 kcal/mol higher in energy, respectively.

Our findings point out that **9n.1** (C_1 , 1A), which is distorted from a high symmetry D_{3h} structure, is the global minimum for neutral Ge_9 (Figure 4). The following isomer is a C_{2v} structure **9n.2** with a relative energy of only 2.4 kcal/mol. Other structures are found to be less stable, being at least 12.3 kcal/mol above.

For Ge_{10} , two degenerate structures **10n.1** (C_{3v} , 1A_1) and **10n.2** (C_1 , 1A), a distorted form of **10n.1**, are found with a tiny energy difference of 0.9 kcal/mol. Similar to the cases of Ge_3 and Ge_8 , this value slightly increases to 1.9 kcal/mol at the CCSD(T)/aug-cc-pVTZ level. Other forms are less stable lying at least 17.0 kcal/mol above. Structure **11n.1** (C_1 , 1A) turns out to be the global minimum for the Ge_{11} system. The nearest isomer is the C_1 structure **11n.2**, which is only 3.1 kcal/mol less stable than the global minimum. For Ge_{12} , our calculated results show that the C_s structure **12n.1** is the lowest isomer. A similar structure with higher symmetry **12n.2**, that is reported to be the most stable isomer at the B3LYP/6-31G(d) level²⁵ is found to have two imaginary frequencies and is much less stable with 17.8 kcal/mol higher in energy. Next isomer is the C_s structure **12n.2** with relative energy of 10.8 kcal/mol.

Low-Lying Isomers of the Monoanions Ge_n^- ($n = 2-12$). Small anionic clusters Ge_n^- ($n = 2-6$) were reported by Xu et al.²¹ and Archibong et al.²² Our calculated results concur well with the previous findings that the structures **2a.1** ($^2\Pi_u$), **3a.1** (C_{2v} , 2A_1), **4a.1** (D_{2h} , $^2B_{2g}$), **5a.1** (D_{3h} , $^1A_2''$), and **6a.1** (D_{4h} , $^2A_{2u}$) (Figure 5) are the global minima of Ge_2^- , Ge_3^- , Ge_4^- , Ge_5^- , and Ge_6^- ,

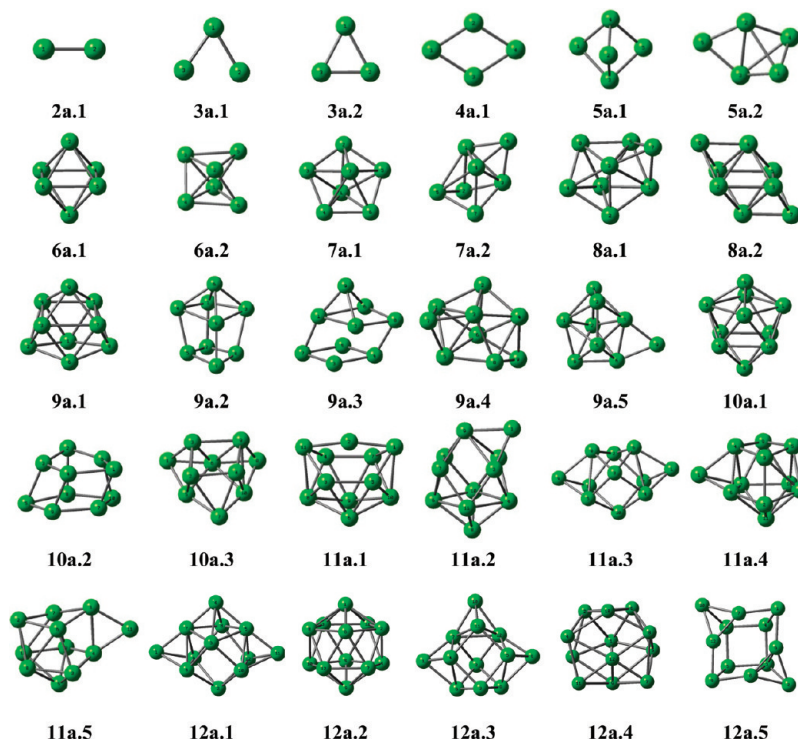


Figure 5. Shapes of the low-lying isomers Ge_n^- ($n = 2-12$).

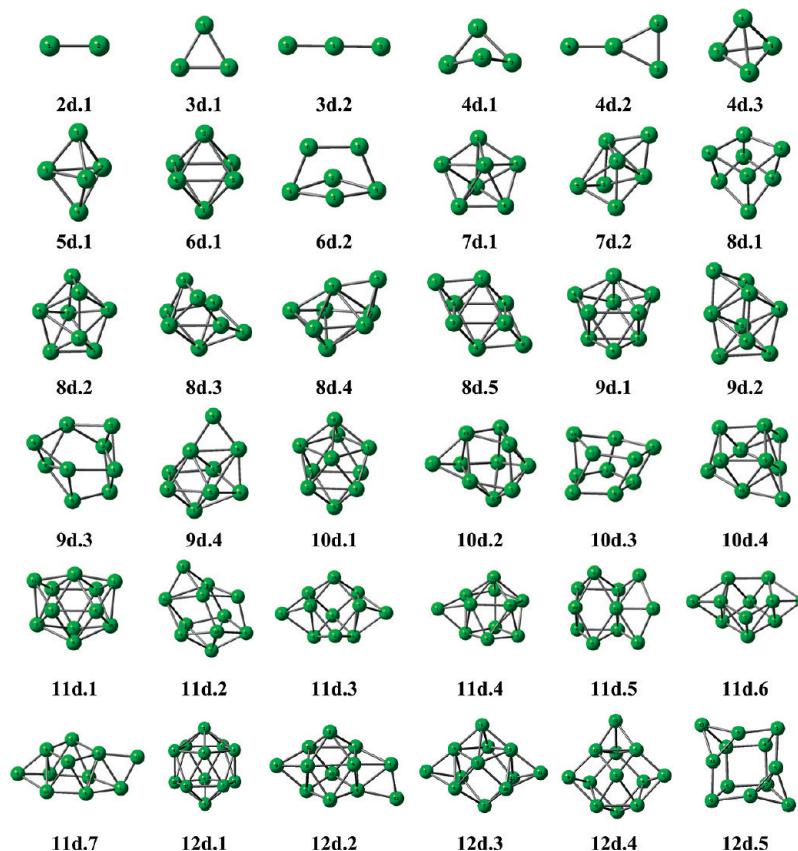


Figure 6. Shapes of the low-lying isomers Ge_n^{2-} ($n = 2-12$).

respectively. Two isomers 3a.1 and 3a.2 with different electronic states are degenerate with energy difference of only 0.1 and 0.5 kcal/

mol at the B3LYP/6-311+G(d) and CCSD(T)/aug-cc-pVTZ levels, respectively.

Table 1. Symmetry Point Group and Electronic State (PG) and Relative Energy (RE, kcal/mol) of the Lower-Lying Neutral Equilibrium Structures Ge_n ($n = 2 - 12$) Using the B3LYP/6-311+G(d) Level

isomers	PG	RE	isomers	PG	RE
2n.1	$D_{\infty h} \ ^3\Sigma_g^-$	0.0	9n.5	$C_s \ ^1A'$	15.7
3n.1	$C_{2v} \ ^1A_1$	0.0	9n.6	$C_1 \ ^1A$	24.0
3n.2	$D_{3h} \ ^3A_1'$	0.5 (1.4) ^a	10n.1	$C_1 \ ^1A$	0.00
3n.3	$D_{\infty h} \ ^3\Sigma_g^+$	10.8	10n.2	$C_{3v} \ ^1A_1$	0.9 (1.9)
4n.1	$D_{2h} \ ^1A_1'$	0.0	10n.3	$C_1 \ ^1A$	18.8
4n.2	$T_d \ ^1A_g$	43.6	10n.4	$C_1 \ ^1A$	17.4
5n.1	$D_{3h} \ ^1A_1'$	0.0	10n.5	$C_{2v} \ ^1A_1$	25.6
5n.2	$C_1 \ ^1A$	11.7	10n.6	$C_{2h} \ ^1A_g$	31.1
5n.3	$C_{2v} \ ^1A_1$	28.9	10n.7	$C_1 \ ^1A$	17.6
5n.4	$C_{2v} \ ^1A_1$	35.3	10n.8	$C_1 \ ^1A$	17.1
6n.1	$C_{2v} \ ^1A_1$	0.0	10n.9	$C_2 \ ^1A$	17.4
6n.2	$C_s \ ^1A'$	0.0 (0.3)	11n.1	$C_1 \ ^1A$	0.0
6n.3	$C_2 \ ^1A$	29.2	11n.2	$C_1 \ ^1A$	3.2
7n.1	$D_{5h} \ ^1A_1'$	0.0	11n.3	$C_1 \ ^1A$	7.4
7n.2	$C_2 \ ^1A$	21.4	11n.4	$C_1 \ ^1A$	9.6
7n.3	$C_s \ ^1A'$	29.1	11n.5	$C_1 \ ^1A$	10.8
7n.4	$C_{3v} \ ^1A_1$	31.4	11n.6	$C_{2h} \ ^1A_g$	11.8
8n.1	$C_{2h} \ ^1A_g$	0.0	11n.7	$C_1 \ ^1A$	13.6
8n.2	$C_s \ ^1A'$	0.7 (1.2)	11n.8	$C_1 \ ^1A$	15.4
8n.3	$C_2 \ ^1A$	3.0	11n.9	$C_{2v} \ ^1A_1$	20.0
8n.4	$C_s \ ^1A'$	4.8	11n.10	$C_s \ ^1A'$	22.3
8n.5	$C_1 \ ^1A$	10.9	12n.1	$C_s \ ^1A'$	0.0
9n.1	$C_1 \ ^1A$	0.0	12n.2	$C_s \ ^1A'$	10.8
9n.2	$C_{2v} \ ^1A_1$	2.4	12n.3	$C_1 \ ^1A$	13.0
9n.3	$C_1 \ ^1A$	12.3	12n.4	$C_{5v} \ ^1A_1$	15.0
9n.4	$C_1 \ ^1A$	14.6	12n.5	$C_{2v} \ ^1A_1$	16.7

^a Values given in parentheses are obtained at the CCSD(T)/aug-cc-pVTZ//B3LYP/6-311+G(d) + ZPE level.

Following attachment of one excess electron into the lowest unoccupied molecular orbital (LUMO) of the neutral Ge_7 , the D_{5h} structure **7a.1** remains the global minimum shape for Ge_7^- . The closest isomer is the C_1 structure **7a.2**, that is 23.8 kcal/mol less stable than **7a.1**. The structure **8a.1**, that is formed by adding one electron to LUMO of the neutral **8n.2**, is indicated to be the lowest isomers for Ge_8^- . The next isomer is the C_{2h} structure **8a.2** with a relative energy of 2.9 kcal/mol.

A number of local minima are located for the anions Ge_9^- , Ge_{10}^- , and Ge_{11}^- . The global minimum of Ge_9^- is the C_1 **9a.1** that is a distortion of the D_{3h} . The structures **10a.1** (C_4) and **11a.1** (C_s) are found to be the lowest-lying isomers of Ge_{10}^- and Ge_{11}^- , respectively. Following addition of one excess electron into the neutral Ge_{12} , the C_s structure **12a.1** remains to be the lowest isomer for anion Ge_{12}^- . However, we also found that the C_{5v} **12a.2** is almost degenerate to the first with relative energy of 0.1 kcal/mol at B3LYP/6-311+G(d) level (0.7 kcal/mol at CCSD(T)/aug-cc-pVTZ level), whereas the C_s **12a.3** is the next isomer with 10.1 kcal/mol higher in energy.

Global Minima of the Dianions Ge_n^{2-} ($n = 2 - 12$). Although relevant experimental studies have not yet carried out, the

Table 2. Point Group and Electronic State (PG) and Relative Energy (RE, kcal/mol) of the Lower-Lying Anionic Equilibrium Structures Ge_n^- and Ge_n^{2-} ($n = 2 - 12$) Using the B3LYP/6-311+G(d) Level

isomers	PG	RE	isomers	PG	RE
2a.1	$D_{\infty h} \ ^2\Pi_u$	0.0	2d.1	$D_{\infty h} \ ^1\Sigma_g^+$	0.0
3a.1	$C_{2v} \ ^2A_1$	0.0	3d.1	$D_{3h} \ ^1A_1'$	0.0
3a.2	$C_{2v} \ ^2B_2$	0.1 (0.5) ^a	3d.2	$D_{\infty h} \ ^1\Sigma_g^+$	22.1
4a.1	$D_{2h} \ ^2B_{2g}$	0.0	4d.1	$D_{2d} \ ^1A_1$	0.0
5a.1	$D_{3h} \ ^2A_2''$	0.0	4d.2	$C_{2v} \ ^1A_1$	15.1
5a.2	$C_1 \ ^2A$	5.0	4d.3	$T_d \ ^3A_1$	20.8
6a.1	$D_{4h} \ ^2A_{2u}$	0.0	5d.1	$D_{3h} \ ^1A_1'$	0.0
6a.2	$C_{2v} \ ^2B_2$	2.1	6d.1	$O_h \ ^1A_{1g}$	0.0
7a.1	$D_{5h} \ ^2A_2''$	0.0	6d.2	$C_{2v} \ ^1A_1$	11.2
7a.2	$C_s \ ^2A'$	23.8	7d.1	$D_{5h} \ ^1A_1'$	0.0
8a.1	$C_s \ ^2A'$	0.0	7d.2	$C_1 \ ^1A$	21.2
8a.2	$C_{2h} \ ^2B_u$	2.9	8d.1	$T_d \ ^1A_1$	0.0
9a.1	$C_1 \ ^2A$	0.0	8d.2	$D_{2d} \ ^1A_1$	3.2
9a.2	$C_1 \ ^2A$	12.0	8d.3	$C_1 \ ^1A$	11.4
9a.3	$C_1 \ ^2A$	13.6	8d.4	$C_s \ ^1A'$	13.3
9a.4	$C_1 \ ^2A$	21.6	8d.5	$C_2 \ ^1A$	14.2
9a.5	$C_1 \ ^2A$	25.0	9d.1	$D_{3h} \ ^1A_1'$	0.0
10a.1	$C_{4v} \ ^2A$	0.0	9d.2	$C_s \ ^1A'$	18.9
10a.2	$C_1 \ ^2A$	19.2	9d.3	$C_1 \ ^1A$	32.1
10a.3	$C_1 \ ^2A$	25.2	9d.4	$C_1 \ ^1A$	36.3
11a.1	$C_s \ ^2A'$	0.0	10d.1	$D_{4d} \ ^1A_1$	0.0
11a.2	$C_1 \ ^2A$	1.6	10d.2	$C_1 \ ^1A$	21.4
11a.3	$C_{2v} \ ^2A_1$	2.1	10d.3	$C_1 \ ^1A$	40.7
11a.4	$C_1 \ ^2A$	2.3	10d.4	$C_1 \ ^1A$	36.7
11a.5	$C_s \ ^2A'$	3.3	11d.1	$C_{2v} \ ^1A_1$	0.0
11a.6	$C_s \ ^2A'$	12.9	11d.2	$C_1 \ ^1A$	3.0
12a.1	$C_s \ ^2A'$	0.0	11d.3	$C_s \ ^1A'$	3.8
12a.2	$C_{5v} \ ^2A_1$	0.1 (0.7)	11d.4	$C_1 \ ^1A$	4.8
12a.3	$C_s \ ^2A''$	10.1	11d.5	$C_1 \ ^1A$	17.9
12a.4	$C_1 \ ^2A$	14.5	11d.6	$C_1 \ ^1A$	24.7
12a.5	$C_{2v} \ ^2A_1$	25.4	12d.3	$C_s \ ^1A'$	10.5
12d.1	$I_h \ ^1A_g$	0.0	12d.4	$C_s \ ^1A'$	22.9
12d.2	$C_s \ ^1A'$	4.5	12d.5	$C_{2v} \ ^1A_1$	36.8

^a Values given in parentheses are obtained at the CCSD(T)/aug-cc-pVTZ//B3LYP/6-311+G(d) + ZPE level.

dianions Ge_n^{2-} are of interest because their structure and stability motif are similar to those of the well-known boron hydrides $\text{B}_n\text{H}_n^{2-}$. While the symmetry point groups and relative energies of the low-lying isomers Ge_n^{2-} are tabulated in Table 2, their shapes are displayed in Figure 6.

The low-spin **2d.1** ($^1\Sigma_g^+$) is found to be the ground state for Ge_2^{2-} with a relatively large singlet–triplet gap of 25.0 kcal/mol. Following addition of two excess electrons into the high-symmetry neutral Ge_3 , the doubly degenerate singly occupied molecular orbitals (SOMOs) of the triplet **3n.2** (D_{3h}) are occupied, and the singlet **3d.1** ($D_{3h} \ ^1A_1'$) becomes consequently the ground state for the dianion Ge_3^{2-} . The D_{2d} structure **4d.1**, that is analogous to the global minimum D_{2d} of boron hydride dianion $\text{B}_4\text{H}_4^{2-}$, is found to be the most stable Ge_4^{2-} isomer.

In agreement with King et al.,^{23,24} the structures **5d.1** (D_{3h}), **6d.1** (O_h), **7d.1** (D_{5h}), and **8d.1** (T_d) are calculated as the global minima for Ge_5^{2-} , Ge_6^{2-} , Ge_7^{2-} , and Ge_8^{2-} , respectively. The

Table 3. Symmetry Point Groups of Ge_n and Ge_n²⁻ (n = 2–12) Clusters in Comparison to the Corresponding Si_n Clusters and Boron Hydrides

Ge _n	Si _n ^a	B _n H _n ^b	Ge _n ²⁻	Si _n ^{2-a}	B _n H _n ^{2-b}
5n.1 (D _{3h})	Si ₅ (D _{3h})	B ₅ H ₅ (C _{4v})	5d.1 (D _{3h})	Si ₅ ²⁻ (D _{3h})	B ₅ H ₅ ²⁻ (D _{3h})
6n.1 (C _{2v})	Si ₆ (C _{2v})	B ₆ H ₆ (C _{2v})	6d.1 (O _h)	Si ₆ ²⁻ (O _h)	B ₆ H ₆ ²⁻ (O _h)
7n.1 (D _{5h})	Si ₇ (D _{5h})	B ₇ H ₇ (C _{3v})	7d.1 (D _{5h})	Si ₇ ²⁻ (D _{5h})	B ₇ H ₇ ²⁻ (D _{5h})
8n.1 (C _{2h})	Si ₈ (C _{2v})	B ₈ H ₈ (D _{2d})	8d.1 (T _d)	Si ₈ ²⁻ (D _{2d})	B ₈ H ₈ ²⁻ (D _{2d})
9n.1 (C ₁)	Si ₉ (C _s)	B ₉ H ₉ (C _{4v})	9d.1 (D _{3h})	Si ₉ ²⁻ (C _s)	B ₉ H ₉ ²⁻ (D _{3d})
10n.1 (C _{3v})	Si ₁₀ (C _{3v})	B ₁₀ H ₁₀ (C _{3v})	10d.1 (D _{4d})	Si ₁₀ ²⁻ (D _{4d})	B ₁₀ H ₁₀ ²⁻ (D _{4d})
11n.1 (C ₁)	Si ₁₁ (C _{2v})	B ₁₁ H ₁₁ (C _{2v})	11d.1 (C _{2v})	Si ₁₁ ²⁻ (C _s)	B ₁₁ H ₁₁ ²⁻ (C _{2v})
12n.1 (C _s)	Si ₁₂ (C _s)	B ₁₂ H ₁₂ (C _{2v})	12d.1 (I _h)	Si ₁₂ ²⁻ (C _{2v})	B ₁₂ H ₁₂ ²⁻ (I _h)

^a Structures silicon clusters are taken from ref 45. ^b Structures for boron hydrides are taken from ref 44.

D_{2d} structure **8d.2** is the next isomer that is 3.2 kcal/mol higher in energy as compared to **8d.1**.

There is indeed an overall analogy between the global minima of Ge_n²⁻ (n = 9–11) and those of the isovalent B_nH_n²⁻. The structure **9d.2** (D_{3h}, ¹A₁') is the global minimum for Ge₉²⁻. Other forms are much less stable being at least 18.9 kcal/mol above the latter. Similarly, the structure **10d.1** (D_{4d}, ¹A₁) turns out to be the global minimum for Ge₁₀²⁻, while the closest C_s isomer **10d.2** is 21.4 kcal/mol higher in energy.

Our calculations point out four low-lying isomers for Ge₁₁²⁻. Although **11d.1** (C_{2v}, ¹A₁) is the global minimum, the other three structures **11d.2**, **11d.3**, and **11d.4** are located with small energy differences.

In agreement with earlier reports,^{17,42} the icosahedral form (I_h) **12d.1** is the global minimum of Ge₁₂²⁻. Next isomer is the C_s **12d.2** with relative energy of 4.5 kcal/mol. While the C_{2v} structure (like **12d.3**) that was reported to be the most stable isomer at the B3LYP/6-31G(d) level²⁵ is only a transition state with two imaginary frequencies, the **12d.3** (C_s) is quite stable with 10.5 kcal/mol above the **12d.1**.

Wade's Rule. The Wade's rule is known as an effective electron count for predicting and interpreting the structure of boron hydrides B_nH_n and hydroborate dianions B_nH_n²⁻.^{43,44} Recently, some studies on the group IVA clusters showed that a structural analogy between the boron hydrides and the isovalent group IVA clusters exists. For instance, Zdetis found a geometrical analogy between silicon clusters Si_n and Si_n²⁻ (n = 5–13) and corresponding deltahedral boranes B_nH_n and B_nH_n²⁻, respectively.⁴⁵ This author detected a strong analogy in their structure at the magic sizes such as Si₆ and Si₁₀. As mentioned above, for the Ge_n^x clusters (n = 5–8 and 12 and x = -4, -2, 0, +2, +4), King and co-workers^{23,24} observed a geometrical analogy between the dianions Ge_n²⁻ with n = 5–7 and boron hydrides B_nH_n²⁻. A question of interest arises is that how far the small germanium clusters Ge_n^x at both neutral and dianion states are analogous to the corresponding boron hydrides.

The symmetry of the lowest-lying isomers of the Ge clusters considered together with those of the Si clusters and boron hydrides are given in Table 3. It can be seen that the molecular structure of the Ge_n neutrals are similar to those of the Si_n neutrals. In the dianionic state, there is also a good analogy between structures of Ge_n²⁻ and B_nH_n²⁻. While the B₈H₈²⁻ dianion has a D_{2d} structure, the dianion Ge₈²⁻ is a T_d structure that is actually a higher connected point group of D_{2d}. Other features are similar between both systems. Also similar to the Si clusters, the analogy to boron hydrides appears to be more marked for Ge₁₀.

Relative Stabilities of Ge_n^x Clusters. The relative stability of the clusters considered can be probed on the basis of the second-order difference in energy (Δ²E) and the average binding energy (BE), that are defined as follows:

$$\Delta^2 E(\text{Ge}_n^x) = E(\text{Ge}_{n-1}^x) + E(\text{Ge}_{n+1}^x) - 2E(\text{Ge}_n^x) \quad (1)$$

$$\text{For the neutrals:} \quad E_b = [nE(\text{Ge}) - E(\text{Ge}_n)]/n \quad (2)$$

$$\text{For the anions:} \quad E_b = [(n-1)E(\text{Ge}) + E(\text{Ge}^-) - E(\text{Ge}_n^-)]/n \quad (3)$$

$$\text{For the dianions:} \quad E_b = [(n-2)E(\text{Ge}) + 2E(\text{Ge}^-) - E(\text{Ge}_n^{2-})]/n \quad (4)$$

The Δ²E value of a Ge_n^x is calculated as the energy difference between two dissociation processes. For example, the Δ²E of neutrals Ge_n can be obtained from: Ge_{n+1} → Ge_n + Ge, and Ge_n → Ge_{n-1} + Ge. As a consequence, it reflects the relative stability of Ge_n^x as compared to that of its two immediate neighbors Ge_{n+1}^x and Ge_{n-1}^x. A high value of Δ²E suggests that the size considered has a high relative stability as compared to its neighbors. The plots of Δ²E displayed in Figure 7 of all systems considered reveal that the remarkably high peaks are found at n = 10 in all neutral, anionic, and dianionic states. This observation is consistent with the previous predictions that the Ge₁₀ neutral is highly stable within the series of small Ge_n clusters. At the neutral and anionic states, the clusters Ge₇^{0/-} are also stable species relative to their neighbors, whereas the dianion Ge₇²⁻ is characterized by a small value of Δ²E.

The plots of average binding energy (E_b) showed in Figure 8 reveal similar trends for the neutrals, anions, and dianions. The E_b values thus tend to increase with the increasing size of clusters, and maximum peaks are again observed at the size of n = 10. It is remarkable that the binding energy of Ge₁₂²⁻ (I_h) is slightly larger as compared to that of Ge₁₁²⁻. Consequently, Ge₁₂²⁻ is expected to be a system with enhanced stability within the series of dianion Ge_n²⁻. Addition of one excess electron to neutrals to form anions increases the average binding energy (E_b) of anions, while dianions show lower values due to their inherent instability.

It is useful to state again that the second-order energy difference given in Figure 7 only suggests about the stability of a cluster with respect to its immediate neighbors. Thus, although both Ge₇ and Ge₁₀ clusters have comparable peaks in this plot,

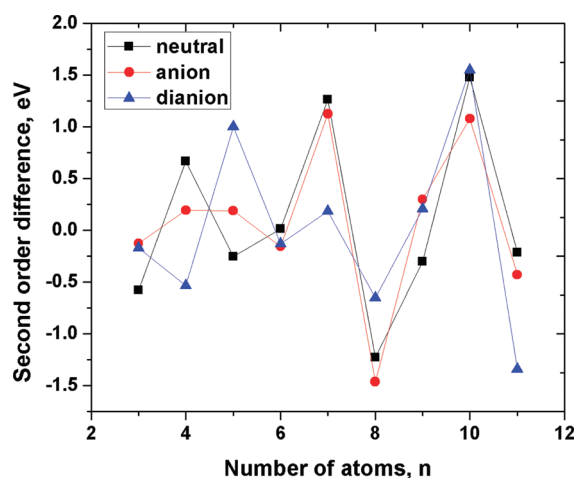


Figure 7. Second-order differences in total energies of germanium clusters Ge_n at neutral, anionic, and dianionic states using the B3LYP/6-311+G(d) level.

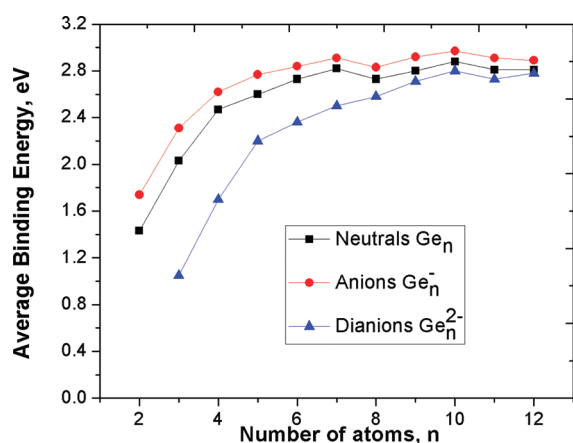


Figure 8. Average binding energies of germanium clusters Ge_n at neutral, anionic, and dianionic states using the B3LYP/6-311+G(d) level.

their relative thermodynamic stability can better be compared from the plot of binding energies illustrated in Figure 8.

Vertical Detachment Energies (VDE) and Adiabatic Detachment Energies (ADE). The values of VDE and ADE can be considered as a measure of stability of an anion with respect to electron removal. For neutrals, these actually correspond to the ionization energies (IE). The VDE corresponds to the difference in energies of the anionic and neutral states at the anion geometry. Similarly, this value of dianion Ge_n^{2-} is obtained by energy difference between the dianion and the anion with geometry of corresponding dianion. The adiabatic values are obtained from energies of the corresponding optimized structures. The VDE and ADE values of the Ge_n , Ge_n^- , and Ge_n^{2-} calculated using both B3LYP and CCSD(T) methods are given in Tables 4–6, respectively. First, it can be observed that there is a negligible difference between the B3LYP and CCSD(T) relative values for small neutrals Ge_n^x ($n = 2-7$). This difference varies in the range of 0.01–0.03 eV for VIE values of neutrals and amounts up to 0.11 eV for VDE's and ADE's of anions. Due to limitation of computational resources, the larger clusters Ge_n^x ($n = 8-11$) are only considered using the B3LYP method, whose results are expected to deviate by ± 0.1 eV relative to the CCSD(T) results.

Table 4. Vertical Ionization Energies (VIE) of the Neutrals Ge_n ($n = 2-12$) Using Both B3LYP/6-311+G(d) and CCSD(T)/aug-cc-pVTZ Levels

neutrals Ge_n	cations Ge_n^+	VIE (eV)		
		B3LYP	CCSD(T)	exptl ^a
2n.1 ($^3\Sigma_g^+$)	Ge_2^+ ($^2\Sigma_g^+$)	8.21	8.22	7.58–7.76
3n.1 (1A_1)	Ge^+ (2B_2)	7.97	7.98	7.97–8.09
4n.1 ($^1A_1'$)	Ge_4^+ ($^2B_{1u}$)	7.83	7.81	7.87–7.97
5n.1 ($^1A_1'$)	Ge_5^+ ($^2E'$)	7.89	7.91	7.87–7.97
6n.1 (1A_1)	Ge_6^+ (2A_1)	7.71	7.74	7.58–7.76
7n.1 ($^1A_1'$)	Ge_7^+ ($^2E_2'$)	7.80		7.58–7.76
8n.1 (1A_g)	Ge_8^+ (2A_u)	6.93		6.72–6.94
9n.1 ($^1A'$)	Ge_9^+ ($^2A'$)	7.24		7.06–7.24
10n.1 (1A)	Ge_{10}^+ (2A)	7.46		7.46–7.76
11n.1 (1A)	Ge_{11}^+ (2A)	6.56		6.55–6.72
12n.1 ($^1A'$)	Ge_{12}^+ ($^2A''$)	6.95		6.94–7.06

^a Experimental values are taken from ref 19.

For the neutrals, our calculated results are in good agreement with the available experimental values.¹⁹ The Ge_7 and Ge_{10} exhibit indeed high VIE values that are followed by drops of VIEs. Consequently, these neutrals are expected to be high stability, with less-stable cationic counterparts, which is consistent with previous reports.^{4,6} In the anionic state, our calculated results for small Ge_n ($n = 2-7$) at both B3LYP and CCSD(T) methods agree well with experimental values and also with previous theoretical predictions. The ADEs of Ge_2^- , Ge_3^- , and Ge_4^- determined at the CCSD(T)/aug-cc-pVTZ +ZPE level amount to 2.05, 2.19, and 2.00 eV, respectively, which can be compared to the corresponding experimental values¹⁹ of 2.035 (Ge_2^-) 2.23 (Ge_3^-), and 1.94 eV (Ge_4^-). For Ge_5^- , the ADE obtained from energy difference between the anion **5a.1** (D_{3h} , $^2A_2''$) and the neutral **5n.1** (D_{3h} , $^1A_1'$) is 2.20 eV, while the ADE calculated from the anion **5a.2** and neutral **5n.2** amounts to 2.49 eV. The latter is in better agreement with the experimental value of 2.51 eV. Similarly, the ADE of 2.04 eV of Ge_6^- agrees well with the experimental value of 2.06 eV, whereas the ADE of Ge_7^- is 1.94 eV by CCSD(T)/aug-cc-pVTZ is slightly larger than the available experimental value of 1.79 ± 0.06 eV.

For Ge_8^- , the ADE evaluated from the energy difference between the anion **8a.1** (C_s , $^2A'$) and the neutral **8n.3** (C_1 , 1A) is equal to 2.36 eV in good agreement with the experimental value of 2.41 eV. Although the pair C_{2h} **8n.1** and **8a3** are stable isomers at both neutral and anionic states, the corresponding ADE value of 2.10 eV obtained from them differs much from the experimental value of 2.41 eV.

Some ADE values obtained from the lowest-lying isomers of larger systems Ge_n^- ($n = 9-11$) are summarized in Table 4. While the ADE value of 2.42 eV of Ge_{11}^- (**11a.2** – $e^- \rightarrow$ **11n.2**) is in a good agreement with experiment (2.50 eV), the deviation between our calculated results and available experimental ADE values of Ge_9^- and Ge_{10}^- is considerably larger. In fact, the ADE of Ge_9^- (**9a.1** – $e^- \rightarrow$ **9n.2**) and Ge_{10}^- (**10a.1** – $e^- \rightarrow$ **10n.1**) is 2.57 and 2.21 eV, respectively, that is consistently but significantly smaller than the corresponding experimental values of 2.86 and 2.50 eV.¹⁸ Two ADE values are calculated for Ge_{12}^- , including (**12a.1** – $e^- \rightarrow$ **12n.1**) and (**12a.2** – $e^- \rightarrow$ **12n.4**). First value of 2.25 eV is in excellent agreement with the experimental value of 2.25 eV,¹⁸ the second of 2.91 eV is much higher due to less stability of the corresponding neutral **12n.4**.

Table 5. Vertical Detachment Energies (VDE) and Adiabatic Detachment Energies (ADE) of the Anions Ge_n^- ($n = 2-12$) Using B3LYP/6-311+G(d) and CCSD(T)/aug-cc-pVTZ Levels

anion	ADE (eV)					VDE (eV)		
	neutral	B3LYP	CCSD(T)	ref 22	exptl. ^a	neutral	B3LYP	CCSD(T)
2a.1 ($^2\Pi_u$)	2n.1 ($^3\Sigma_g^+$)	1.94	2.05	1.95	2.035 ± 0.001	Ge_2 ($^3\Sigma_g^+$)	2.01	2.07
3a.1 (2A_1)	3n.1 (1A_1)	2.18	2.19	2.15	2.23 ± 0.01	Ge_3 (1A_1)	2.49	2.42
4a.1 ($^2B_{2g}$)	4n.1 ($^1A_1'$)	1.94	2.00	1.80	1.94	Ge_4 (1A_g)	1.96	2.02
5a.1 ($^2A_2''$)	5n.1 ($^1A_1'$)	2.20	2.30	2.10		Ge_5 ($^1A_1'$)	2.87	2.84
5a.2 (2A)	5n.2 (1A)	2.49	2.54	—	2.51	Ge_5 (1A)	3.29	3.35
6a.1 ($^2A_{2u}$)	6n.1 (1A_1)	2.01	2.04	2.01	2.06	Ge_6 (1A_1)	2.60	2.61
7a.1 ($^2A_2''$)	7n.1 ($^1A_1'$)	1.99	1.94	—	1.80	Ge_7 ($^1A_1'$)	2.30	2.30
8a.1 ($^2A'$)	8n.3 (1A)	2.36				Ge_8 ($^1A'$)	2.70	
8a.2 (2B_u)	8n.1 (1A_g)	2.10			2.41	Ge_8 ($^1A'$)	2.44	
9a.1 ($^2A'$)	9n.2 (1A_1)	2.57			2.86	Ge_9 ($^1A'$)	3.20	
10a.1 (2A_1)	10n.1 (1A)	2.21			2.50	Ge_{10} (1A_1)	3.16	
11a.1 ($^2A'$)	11n.9 (1A)	3.20				Ge_{11} ($^1A'$)	2.93	
11a.2 ($^2A'$)	11n.2 (1A)	2.42			2.50	Ge_{11} ($^1A'$)	2.83	
11a.3 ($^2A'$)	11n.1 (1A)	2.27				Ge_{11} ($^1A'$)	2.96	
12a.1 ($^2A'$)	12n.1 ($^1A'$)	2.25			2.25	Ge_{12} ($^1A'$)	2.91	
12a.2 (2A_1)	12n.4 (1A_1)	2.91				Ge_{12} (1A_1)	3.43	

^a Experimental values are taken from ref 18.

Table 6. Vertical (VDE) and Adiabatic (ADE) Detachment Energies of the Dianions Ge_n^{2-} ($n = 2-12$) Using B3LYP/6-311+G(d) Level

dianions	ADE		VDE	
	anions	B3LYP	anions	B3LYP
2d.1 ($^1\Sigma_g^+$)	2a.1 ($^2\Pi_u$)	-2.47	Ge_2^- ($^2\Pi_u$)	-2.46
3d.1 ($^1A_1'$)	3a.1 (2A_1)	-2.45	Ge_3^- ($^2E'$)	-2.36
4d.1 (1A_1)	4a.1 ($^2B_{2g}$)	-2.36	Ge_4^- (2A_1)	-1.81
5d.1 ($^1A_1'$)	5a.1 ($^2A_2''$)	-1.56	Ge_5^- ($^2A_2''$)	-1.23
6d.1 ($^1A_{1g}$)	6a.1 ($^2A_{2u}$)	-1.56	Ge_6^- ($^2A_{1g}$)	-1.18
7d.1 ($^1A_1'$)	7a.1 ($^2A_2''$)	-1.60	Ge_7^- ($^2A_2''$)	-1.25
8d.1 (1A_1)	8a.1 (C_{1s})	-0.81	Ge_8^- (2A_1)	-0.62
9d.1 ($^1A_1'$)	9a.1 ($^2A'$)	-0.59	Ge_9^- ($^2A_2''$)	-0.45
10d.1 (1A_1)	10a.1 (2A)	-0.39	Ge_{10}^- (2A_1)	-0.16
11d.1 (1A_1)	11a.1 ($^2A'$)	-0.67	Ge_{11}^- (2A_1)	-0.47
12d.1 (1A_g)	12a.2 (2A_1)	-0.03	Ge_{12}^- (2E_g)	0.27

Most of the free dianions Ge_n^{2-} are not stable with respect to electron detachment, as indicated by their negative VDE and ADE values. The Ge_{12}^{2-} shows a positive VDE value of 0.37 eV, which lends a further support for its enhanced stability.

HOMO–LUMO Gaps. The frontier orbital energy gaps can also be regarded as a measure of kinetic stability. A large gap suggests a relatively low reactivity. The gaps of Ge_n obtained at the B3LYP/6-311+G(d) level are summarized in Table 7. Similar to the trends observed from the second order difference in energy Δ^2E and VDE (ADE), the highest occupied molecular orbital–lowest unoccupied molecular orbital (HOMO–LUMO) gaps are found to be high at the size $n = 10$, that again points out an enhanced stability of Ge_{10}^x within the Ge_n series. As observed from BE and VDE values, the HOMO–LUMO gap of Ge_{12}^{2-} is equal to 3.0 eV, which corresponds to the highest value within the series of dianions. Although Ge_7 also has a local

Table 7. HOMO–LUMO Gaps (HLG, eV) of the Global Minima Ge_n^x ($n = 2-12$, $x = 0, -2$)^a

isomer	HLG	isomer	HLG
2n.1	0.81	2d.1	1.17
3n.1	2.33	3d.1	1.55
4n.1	2.35	4d.1	1.84
5n.1	3.06	5d.1	2.27
6n.1	3.02	6d.1	2.22
7n.1	2.86	7d.1	1.95
8n.1	2.27	8d.1	2.59
9n.1	2.66	9d.1	2.06
10n.1	3.03	10d.1	2.92
11n.1	2.18	11d.1	2.11
12n.1	2.84	12d.1	3.05

^a At the B3LYP/6-311+G(d) level.

maximum peak on the plot of the second-order difference (see above), its HOMO–LUMO gap of 2.8 eV is slightly smaller than that of 3.0 eV for Ge_{10} .

Enhanced Stability and JSM. Finding a consistent rationalization for stability of clusters is always an important goal. In previous reports, stability of the Ge_n clusters can be interpreted using a simple shell model which assumes that each Ge atom distributes two valence electrons from its p-subshell into the electron shell configuration of system. As a result, Ge_{10} contains 20 valence electrons and becomes a thermodynamically stable species due to a possession of a filled electron shell configuration, similar to the situation of alkali metals.

In order to probe the features related to the stability of Ge_n clusters, we re-examine their MO pictures under the viewpoint of the JSM,⁴⁰ which is applied successfully to interpret the stability motif of different types of atomic clusters. According to this rather simple model in which the valence electrons are assumed to be freely itinerant in a simple mean-field potential formed by

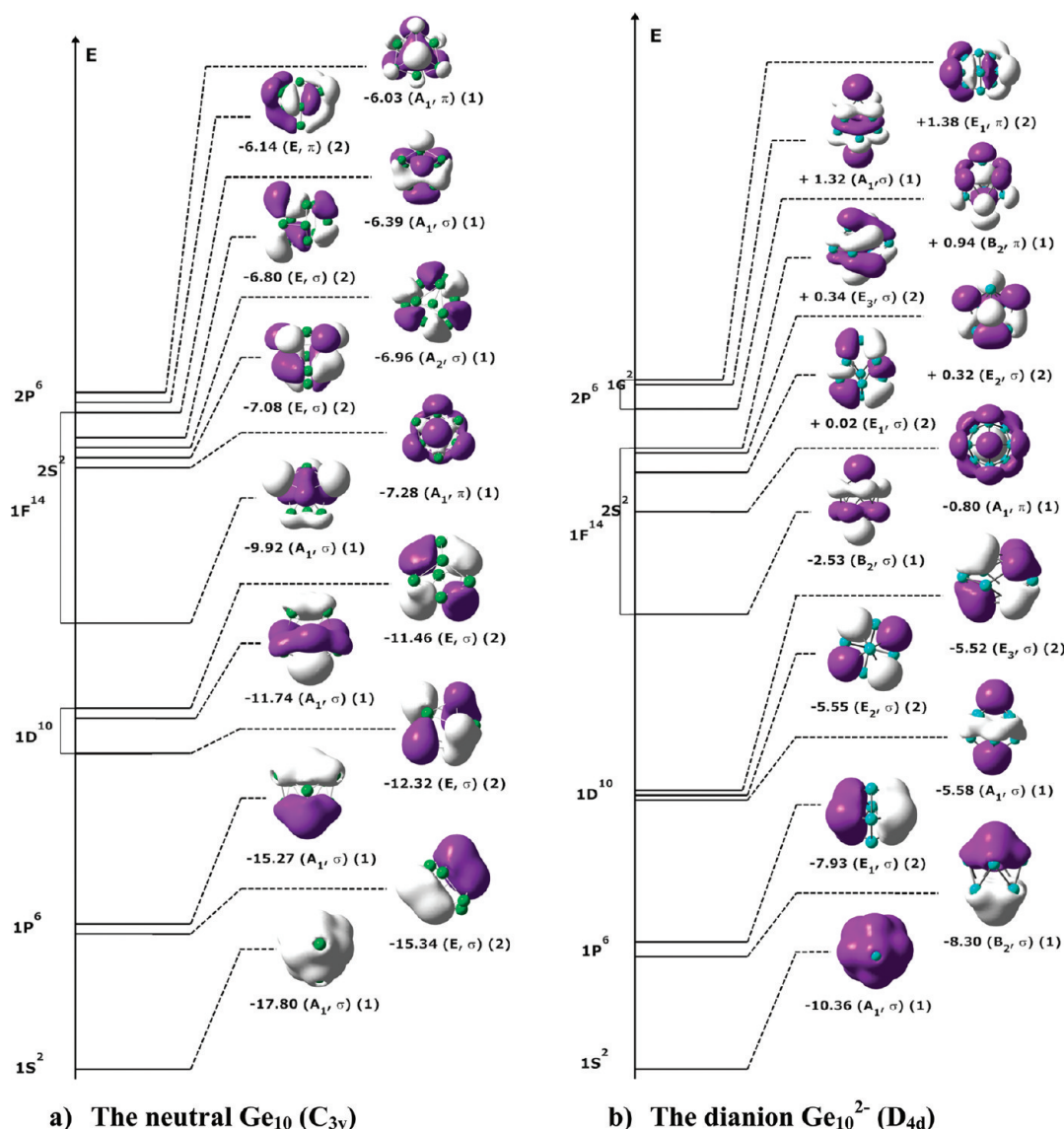


Figure 9. Energy levels and orbitals of the Ge_{10} (C_{3v}) and Ge_{10}^{2-} (D_{4d}) using the B3LYP/6-311+G(d) level.

the nuclei of atoms, the valence electrons fill the spherical orbitals of a system according to the pattern of $[1S^2 1P^6 1D^{10} 2S^2 1F^{14} 2P^6 1G^{18} 2D^{10} \dots]$, etc. As a consequence, the number of electrons of 8, 20, 34, 40, 56, and 68, etc. are proposed as the magic numbers that correspond to the closed shell electrons. We should note that the observed number of electrons of a simple JSM is predicted on the basis of a spherical background. Thus, these magic numbers can be changed due to a lowering of the molecular symmetry

As for a typical case, let us show the MO pictures of Ge_{10} that features an enhanced stability in the neutral state. Each Ge atom is expected to contribute its four valence electrons to the electron shell configuration of the system. The Ge_{10} cluster thus contains 40 valence electrons with an orbital configuration of $[1a_1^2 1e^4 2a_1^2 2e^4 3a_1^2 3e^4 4a_1^4 4e^4 1a_2^2 5e^4 5a_1^2 6e^4 6a_1^2]$ that can be arranged into the model energy ordering of $[1S^2 2P^6 2S^2 1D^{10} 1F^{14} 2P^6]$ (Figure 9). While the HOMO of the Ge_{10} is a p-orbital, its LUMO is a g-orbital, belongs to the model G-subshell.

Due to a lowering from a spherical background to the C_{3v} point group symmetry, the model subshells are split into different

energy levels. However, the resulting energy ordering of Ge_{10} is consistent with the electron shell configuration of 40 electrons of JSM. Overall, the Ge_{10} is a magic size within the Ge_n series.

Adding two excess electrons into the HOMO of the neutral Ge_{10} leads to an unstable system with open electronic shell configuration. However, due to a distortion from the spherical symmetry to D_{4d} symmetry, the model G-subshell of Ge_{10}^{2-} is split into P-subshells (Figure 9b). The valence electrons of the Ge_{10}^{2-} are thus composed of an orbital configuration of $[1a_1^2 1b_2^2 1e_1^4 2a_1^2 1e_2^4 1e_3^4 2b_2^2 3a_1^2 2e_1^4 2e_2^4 2e_3^4 3b_2^2 4a_1^2 3e_1^4]$, which corresponds to the model energy ordering of $[1S^2 1P^6 1D^{10} 2S^2 1F^{14} 2P^4 1G^2 2P^2]$. Its HOMO is the third orbital of 2P-subshell, while its LUMO belongs to a G-subshell. This gives rise to a consequence that Ge_{10}^{2-} possesses a closed shell configuration and thereby a higher stability, as compared to other Ge_n^{2-} dianions.

The enhanced stability of the icosahedral Ge_{12}^{2-} (I_h) is even more interesting. The Ge_{12}^{2-} 12d.1 contains thus 50 valence electrons that are arranged into an energy ordering as $[1S^2 1P^6 1D^{10} 1F^6 2S^2 1F^8 2P^6 1G^{10}]$ (Figure 10). Similar to systems containing 50 valence electrons reported previously, including

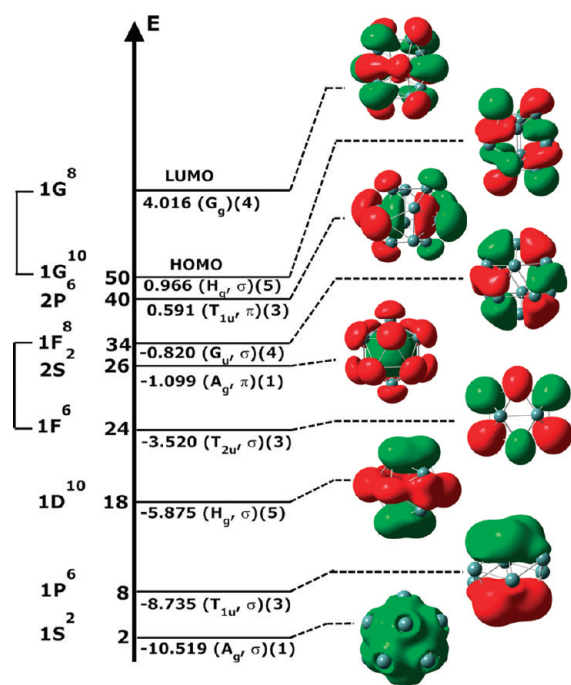


Figure 10. Energy levels and orbitals of the Ge_{12}^{2-} (I_h) using the B3LYP/6-311+G(d) level.

$\text{Pb}_{12}\text{Al}^{+46}$, Ge_{12}M^x ($\text{M} = \text{Li}, \text{Na}, \text{Be}, \text{Mg}, \text{B}, \text{and Al}$),¹⁷ a splitting of the $l = 4$ shell (G-shell) in the icosahedral $12d.1$ (I_h) occurs under the crystal field effects.⁴⁷ This splitting of G-shell consequently results in a large energy gap of 3.0 eV between the frontier orbitals that ultimately leads to its high stability.

Spherical Aromaticity. Aromaticity is usually considered as one of the important measures to probe the thermodynamic stability of chemical compounds. While the Hückel rule of $(4n + 2)$ electrons is popularly applied to determine the aromatic character of planar structures, the I_h symmetrical fullerenes were reported to have spherical aromaticity through the $2(N + 1)^2$ rule, that was recently proposed by Hirsch et al.⁴⁸ The π -electron system of these species can approximately be considered as a spherical electron gas, which surrounds the surface of a sphere. The wave functions of this electron gas can be characterized by the angular momentum quantum numbers ($l = 0, 1, 2, 3, \dots$), that are comparable to the atomic s, p, d, f, ... orbitals. According to the classical Pauli principle, if a system with $2(N + 1)^2$ π -electrons fully fills all π -shells, then it then exhibits an aromatic character.

The MO picture of Ge_{10} (Figure 9a) reveals two orbital sets: While the first contains the σ -type MOs of 1S, 1P, 1D, and 1F, that are occupied by 32 σ -electrons, the second set includes the MOs of 2S (A_1, π) and 2P (E, π and A_1, π), that are thus occupied by 8 valence π -electrons. Consequently, the Ge_{10} system is characterized by eight π -electrons, that make it spherically aromatic, which is consistent with the $2(N + 1)^2$ rule.

Similar systems of eight valence π -electrons are also found for the global minima Ge_{10}^{2-} (D_{4d}) and Ge_{12}^{2-} (I_h) (Figures 9b and 10, respectively) in which each includes two π -electrons belonging to the 2S-subshell and six π -electrons belonging to the 2P-subshell. As a consequence, these dianions Ge_{10}^{2-} and Ge_{12}^{2-} possess an aromatic character that is in line with their enhanced thermodynamic stability. Ge_{10} is found to have an enhanced stability, in good agreement with previous studies. Wang et al.⁴ showed a maximum peak at Ge_{10} in their plot of fragmentation

energies, while the maximum ionization potential of the Ge_n clusters was found at $n = 10$.¹⁰ Thus, the spherical aromaticity is proposed to evaluate the aromaticity of the fullerene-like structures and is not applied to structures like Ge_7 . The stability of Ge_7 is not due to the effect of spherical aromaticity.

4. CONCLUDING REMARKS

In this theoretical study, we carry out a search for the energetically lower-lying isomers of small germanium clusters and the anions and dianions, Ge_n^x with $n = 2-12$ and $x = 0, -1, -2$ using a stochastic method. An improved search method for structures is implemented, and the obtained results are compared to previous reports. Using the new procedure with additional control parameters, optimization yield raises up 90%, and the larger number of isomers located shows the efficiency of the procedure. The structures of global minima Ge_n^x ($x = 0, -2$) are found to be analogous to those of boron hydrides B_nH_n^x . The negatively charged clusters $\text{Ge}_n^{-/2-}$ are systematically characterized, and the energetic results obtained using both B3LYP and CCSD(T) methods are in good agreement with available experimental values and also point out some disagreements. Calculated results show that the Ge_{10}^x clusters, in neutral and dianionic states, and Ge_{12}^{2-} clusters are the magic species with large HOMO–LUMO gaps, high VDE and ADE values, and average binding energies. The enhanced stability of Ge_{10} , Ge_{10}^{2-} , and Ge_{12}^{2-} can consistently be rationalized by using the jellium electron shell model and their high spherically aromatic character.

■ ASSOCIATED CONTENT

S Supporting Information. Tables list the total electronic energies, zero-point energies of the low-lying isomers. Figures display all isomers located for the Ge_n clusters. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: minh.nguyen@chem.kuleuven.be.

■ ACKNOWLEDGMENT

The authors are indebted to the K.U. Leuven Research Council for continuing support (GOA, IDO, and IUAP programs). T.B.T. thanks the Arenberg Doctoral School for a scholarship. M.T.N. thanks the ICST of Ho Chi Minh City for supporting his stays in Vietnam.

■ REFERENCES

- Li, C.; John, S.; Banerjee, S. *J. Electron. Mater.* **1995**, *24*, 875.
- Kohl, V. G. *Z. Naturforsch.* **1954**, *9A*, 913.
- Kikuchi, E.; Ishii, S.; Ohno, K. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2006**, *74*, 195410.
- Wang, J.; Han, J. G. *J. Chem. Phys.* **2005**, *123*, 244303.
- Li, B. X.; Cao, P. L. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2000**, *62*, 15788.
- Bandyopadhyay, D.; Sen, P. J. *Phys. Chem. A* **2010**, *114*, 1835.
- Shvartsburg, A. A.; Liu, B.; Lu, Z. Y.; Wang, C. Z.; Jarrold, M. F.; Ho, K. M. *Phys. Rev. Lett.* **1999**, *83*, 2167.
- Wang, J.; Wang, G.; Zhao, J. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2001**, *64*, 205411.

- (9) (a) Yoo, S.; Zeng, X. C. *J. Chem. Phys.* **2006**, *124*, 184309. (b) Bulusu, S.; Yoo, S.; Zeng, Z. C. *J. Chem. Phys.* **2005**, *122*, 164305.
- (10) Yoshida, S.; Fuke, K. *J. Chem. Phys.* **1999**, *111*, 3880.
- (11) Jarrold, M. F.; Bower, J. E. *J. Chem. Phys.* **1992**, *96*, 9180.
- (12) Gopakumar, G.; Lievens, P.; Nguyen, M. T. *J. Chem. Phys.* **2006**, *124*, 214312.
- (13) Gopakumar, G.; Lievens, P.; Nguyen, M. T. *J. Phys. Chem. A* **2007**, *111*, 4355.
- (14) Hou, X. J.; Gopakumar, G.; Lievens, P.; Nguyen, M. T. *J. Phys. Chem. A* **2007**, *111*, 13544.
- (15) Gopakumar, G.; Ngan, V. T.; Lievens, P.; Nguyen, M. T. *J. Phys. Chem. A* **2008**, *112*, 12187.
- (16) Gopakumar, G.; Wang, X.; Lin, L.; De Haeck, J.; Lievens, P.; Nguyen, M. T. *J. Phys. Chem. C* **2009**, *113*, 10856.
- (17) (a) Tai, T. B.; Nguyen, M. T. *Chem. Phys. Lett.* **2010**, *492*, 290. (b) Holtz, T.; Veldeman, N.; Veszpremi, T.; Lievens, P.; Nguyen, M. T. *Chem. Phys. Lett.* **2009**, *469*, 304.
- (18) (a) Burton, G. R.; Xu, C.; Arnold, C. C.; Neumark, D. M. *J. Chem. Phys.* **1996**, *104*, 2757. (b) Burton, G. R.; Xu, C.; Neumark, D. M. *Surf. Rev. Lett.* **1996**, *3*, 383.
- (19) Yoshida, S.; Fuke, K. *J. Chem. Phys.* **1999**, *111*, 3880.
- (20) Deutsch, P. W.; Curtiss, L. A.; Blaudeau, J. P. *Chem. Phys. Lett.* **1997**, *270*, 413.
- (21) Xu, W.; Zhao, Y.; Li, Q.; Xie, Y.; Schaefer, H. F., III. *Mol. Phys.* **2004**, *102*, 579.
- (22) Archibong, E.; St-Amant, A. *J. Chem. Phys.* **1998**, *109*, 962.
- (23) King, R. B.; Dumitrescu, I. S.; Kun, A. *J. Chem. Soc., Dalton Trans.* **2002**, 3999.
- (24) King, R. B.; Dumitrescu, I. S.; Lupan, A. *J. Chem. Soc., Dalton Trans.* **2005**, 1858.
- (25) King, R. B.; Dumitrescu, I. S.; Uta, M. M. *J. Chem. Soc., Dalton Trans.* **2007**, 364.
- (26) Nguyen, M. T.; Matus, M. H.; Dixon, D. A. *Inorg. Chem.* **2007**, *46*, 7561.
- (27) Tai, T. B.; Nguyen, M. T. *Kwant-Kick*: A program for searching cluster structures using a stochastic approach, K. U. Leuven (2010)
- (28) Zhao, J.; Xie, R. *J. Comput. Theor. Nanosci.* **2004**, *1*, 117.
- (29) Alexandrova, A. N.; Boldyrev, A. I. *J. Chem. Theory Comput.* **2005**, *1*, 566.
- (30) Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471.
- (31) Goedecker, S. *J. Chem. Phys.* **2004**, *120*, 9911.
- (32) Saunders, M. *J. Comput. Chem.* **2003**, *25*, 621.
- (33) Bera, P. P.; Sattelmeyer, K. W.; Saunders, M.; Schaefer, H. F., III; Schleyer, P. v. R. *J. Phys. Chem. A* **2006**, *110*, 4287.
- (34) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, Jr., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; and Pople, J. A. *Gaussian 03*, revision C.01; Gaussian, Inc.: Wallingford, CT, 2004.
- (35) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785.
- (36) Rassolov, V. A.; Ratner, M. A.; Pople, J. A.; Redfern, P. C.; Curtiss, L. A. *J. Comput. Chem.* **2001**, *22*, 976.
- (37) Binning, R. C.; Curtiss, L. A. *J. Comput. Chem.* **1990**, *11*, 1206.
- (38) Pople, J. A.; Head-Gordon, M.; Raghavachari, K. *J. Chem. Phys.* **1987**, *87*, 5968.
- (39) Wilson, A. K.; Woon, D. E.; Perterson, K. A.; Dunning, T. H. *J. Chem. Phys.* **1999**, *110*, 7667.
- (40) Brack, M. *Rev. Mod. Phys.* **1993**, *65*, 677.
- (41) (a) Tai, T. B.; Nhat, P. V.; Nguyen, M. T. *Phys. Chem. Chem. Phys.* **2010**, *12*, 11477. (b) Tai, T. B.; Nguyen, M. T. *Chem. Phys. Lett.* **2010**, *489*, 75. (c) Holtz, T.; Veldeman, N.; De Haeck, J.; Veszpremi, T.; Lievens, P.; Nguyen, M. T. *Chem.—Eur. J.* **2009**, *15*, 3970.
- (42) Shao, N.; Bulusu, S.; Zeng, X. C. *J. Chem. Phys.* **2008**, *128*, 154326.
- (43) Kalvoda, S.; Paulus, B.; Dolg, M.; Stoll, H.; Werner, H. J. *Phys. Chem. Chem. Phys.* **2001**, *3*, 514.
- (44) McKee, M. L.; Wang, Z. X.; Schleyer, P. v. R. *J. Am. Chem. Soc.* **2000**, *122*, 4781.
- (45) Zdetsis, A. D. *J. Chem. Phys.* **2007**, *127*, 244308.
- (46) Neukermans; Janssens, K.; Chen, Z. F.; Silverans, R. E.; Schleyer, P. v. R.; Lievens, P. *Phys. Rev. Lett.* **2004**, *92*, 163401.
- (47) Schriver, K. E.; Persson, J. L.; Honea, E. C.; Whetten, R. L. *Phys. Rev. Lett.* **1990**, *64*, 2539.
- (48) Hirsch, A.; Chen, Z.; Jiao, H. *Angew. Chem., Int. Ed.* **2000**, *39*, 3915.

Molecular Dynamics Simulation Study of Chlorophyll a in Different Organic Solvents

Khadga Karki and Danilo Roccatano*

School of Engineering and Science, Jacobs University Bremen, Campus Ring 1, D-28759, Bremen, Germany

S Supporting Information

ABSTRACT: Herein, we present a new model of chlorophyll a for molecular dynamics simulations based on the optimized potentials for liquid simulations force field. The new model was used to study the structural and dynamic properties of the molecule in three different solvents: water, methanol, and benzene. The results of the simulations show that structural and dynamic properties of the chlorin ring are similar in both methanol and benzene. In methanol and water, the magnesium in the chlorin ring binds the oxygen of the solvent molecules with residence times of 2566 and 1300 ps, respectively. In both methanol and benzene, the phytol tail shows a worm-like chain distribution with a larger persistence length for the molecule in benzene. On the contrary, chlorophyll a in water adopts a more compact structure with the phytol chain folded onto the chlorin ring. This conformation is consistent with the expected conformation of the aggregates of chlorophyll a in aqueous environments. Finally, the rotational time constants obtained with our model from the simulations in methanol (125 ps) and benzene (192 ps) are in good agreement with the value extrapolated from the experimental data.

INTRODUCTION

Chlorophylls are among the most important molecules in nature.¹ Chlorophyll a (Chl A), one of the chlorophyll molecules, plays a key role in the light-harvesting complex (LHC), by collecting and funneling light, and as an electron carrier in the photosynthetic reaction centers, by separating charges and transferring electrons across the photosynthetic membrane.^{2,3} These marvelous biological processes are the result of cooperative effects, depending crucially not only on the electronic properties of an individual Chl A molecule but also on the way they are assembled.⁴ The structural organization of the chlorophyll molecules in photosynthetic systems is orchestrated by electrostatic and van der Waals and other nuanced interactions of different functional groups with the surrounding environment. The phytol tail of Chl A is hydrophobic, while the magnesium (Mg), being coordinatively unsaturated, attracts nucleophilic polar molecules. In nonpolar solvents, like benzene, the coordination of Mg is saturated by the electron donor C=O group of another Chl A molecule leading to the formation of dimers and aggregates. In polar solvents, like methanol, the nucleophilic solvent molecules compete with the C=O group for the coordination with the Mg thereby preventing aggregation. However, in other polar solvents, like water, they form large aggregates² because of the intermolecular hydrogen bonding or smaller aggregates, like in mixture of acetonitrile/water, because of the hydrophobic effects of the phytol tail.⁵

Most of the experimental techniques employed to understand these interactions (for example see the references)^{6–11} do not provide the information with atomic resolution. Therefore, use of theoretical/computational model can complement these data and provides useful insights for interpretation and comprehension of the interactions. Several computational studies, based on molecular dynamics (MD) simulations and quantum mechanics (QM) methods, on the structural and spectroscopic properties of chlorophylls are available in literature.¹² MD simulation studies

of chlorophylls, including the ones embedded in the LHCs (see for example refs 13–16), have also been reported. However, to the best of our knowledge, none of these models have been optimized and tested against the properties of isolated molecules in different solvents. In this first paper, we use a new model of Chl A based on the optimized potentials for liquid simulations (OPLS) force field to study the structural and dynamic properties of Chl A in three different solvents: water, methanol, and benzene. We have considered methanol (dielectric constant $\epsilon = 33$) and benzene ($\epsilon = 2.28$) because both of them dissolve Chl A, and hence, it is interesting to analyze their effect on the conformation of the molecule in the different environments. On the contrary, simulations in water can be used to test the force field and to get insights into the mechanism of water-mediated aggregation. In the paper, we have focused mainly on investigating the interaction of Chl A with the solvent molecules by comparing our results with the experimental data¹⁷ and recently published QM calculations.¹⁸

The paper is organized as follows. The modeling of the Chl A force field is presented in the Material and Methods Section. The structural and dynamic properties of the molecule are presented in the Results and Discussion Section. In this part, we also report a preliminary study on the aggregation of Chl A. Finally, in the Conclusion Section, a summary of the study with an outlook on the further applications of the model is presented.

MATERIAL AND METHODS

QM Calculations. Chlorophyllide a molecule was used instead of Chl A to calculate the binding energy of water and methanol to the Mg. The starting structure of chlorophyllide a with a water molecule coordinated to the Mg was obtained from the crystal

Received: August 18, 2010

Published: February 24, 2011

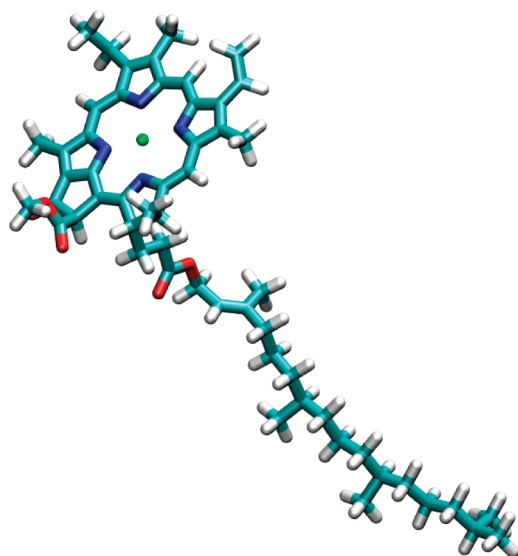


Figure 1. Optimized geometry of Chlorophyll a in vacuum.

structure of the ethyl derivative of the molecule.¹⁹ The molecule was isolated from water, and the ethyl group was substituted by a hydrogen to retain the chlorophyllide a structure. Geometry optimization was done with B3LYP method and 6-31G** basis set in implicit solvation condition using the integral–equation–formalism protocol²⁰ formulation of the polarizable continuum model (PCM).²¹ A water molecule was placed 0.3 nm above the Mg of the optimized chlorophyllide, and the geometry of the complex was optimized again using the same procedure, after which counterpoise correction was used to calculate the binding energy of the water molecule to the Mg. Normal mode analysis was done on the water and the chlorophyllide molecules separately to estimate the thermal corrections used to calculate the Helmholtz free energy. Similar calculations were done in the case of methanol and benzene. The optimized geometries used in the QM calculations are given in the Supporting Information.

QM calculations were also done on Chl A taken from the crystal structure of the LHC of spinach²² (pdb code: 1RWT) to obtain the optimized structure (shown in Figure 1) and the partial charges of the molecule. Geometry optimization was done using the restricted B3LYP method with the 6-31G* basis set, and the atomic charges were calculated by fitting the molecular electrostatic potential of the electronic density using the CHELPG procedure.²³ The coordinates of the optimized structure and the corresponding partial charges are reported in Table 1 and 2 of the Supporting Information.

All the QM calculations were performed using the program Gaussian 09.²⁴

Chl A Force Field Parameters. We used the OPLS-AA^{25–28} force field with additional parameters for partial charges, bond lengths, angles, and dihedral angles based upon our QM calculations. The force constants for the bond angles and the torsional interactions and the Lennard-Jones parameters were adapted from the OPLS parameters, while the partial charges were taken from the QM calculations described above. The full set of parameters are reported in the Supporting Information. OPLS force field-based models were also used for methanol,²⁵ water (TIP4P model),²⁹ and benzene.³⁰

MD Simulations. All MD simulations and analysis were performed using the GROMACS package (version 4.0.7).³¹

The geometry optimized Chl A molecule was immersed in three boxes containing the three different solvents. The simulations in water were done in a 5.34 nm long cubic box containing 4999 water molecules coupled to an external bath at 300 K using the Nose–Hoover³² thermostat and to a barostat at 1 bar using the Parrinello–Rahman^{33,34} isotropic pressure coupling method, as implemented in GROMACS. The coupling time constants for the control of the temperature (τ_T) and the pressure (τ_P) were set to 0.2 and 1.0 ps, respectively, and the compressibility was set to $4.5 \times 10^{-5} \text{ bar}^{-1}$. For the simulations in methanol, a cubic box of length 4.9 nm containing 1685 molecules was used. The coupling time constants τ_T and τ_P were set to 0.1 and 1.5 ps, respectively, and the compressibility was set to $1.2 \times 10^{-4} \text{ bar}^{-1}$. For the simulations in benzene, a cubic box of length 6.31 nm containing 1685 solvent molecules was used. The coupling time constants τ_T and τ_P were set to 0.2 and 2.5 ps, respectively, and the compressibility was set to $9.5 \times 10^{-5} \text{ bar}^{-1}$. All the solvent molecules within 0.15 nm of any Chl A atom were removed, and the systems were energy minimized with the steepest-descent method for 5000 steps. The bond lengths were constrained using the SETTLE algorithm³⁵ for the water molecules and the LINCS algorithm³⁶ for the other molecules. A 1.0–1.2 nm switched cutoff radius was used for the Lennard-Jones interactions. The PME method³⁷ was used for the electrostatic interactions with PME order of 4, Fourier spacing of 0.12 nm, and dielectric permittivity of 1. The short-range neighbor list was set to 1.4 nm. All the atoms were given an initial velocity obtained from a Maxwellian distribution at 300 K. A time step of 2 fs was used in the simulations, and they were equilibrated by 500 ps of MD runs to allow the relaxation of the solvent molecules. After the equilibration, 50 ns production run was performed for each simulation. Simulations were also performed at constant temperature and volume (NVT) conditions to calculate the potential of mean force (PMF)³⁸ of the interaction of the solvent molecules with the Mg. Simulations of systems with 10 Chl A molecules in the three solvents were also performed to test the formation of aggregates in the different solvents. The boxes used in the simulations contained 10 889, 10 828, and 10 827 molecules of water, methanol, and benzene, respectively.

Cluster Analysis of the Chl A Conformation. A reliable estimation of the conformational space explored by the simulations is the evaluation of the number of different configurations generated during the trajectory.³⁹ The cluster analysis of trajectories was performed using the method proposed by Daura et al.⁴⁰ on a total of 5000 structures sampled every 10 ps. The clustering algorithm was applied to the heavy atoms of Chl A. The criteria of similarity for two structures was a positional root-mean-square deviation with the cutoff set to 0.3 nm. Similar analysis was done to the chlorin ring using the cutoff of 0.02 nm.

Phytol Tail analysis. The conformational dynamics of the phytol chain was analyzed by calculating the distribution of the beginning-to-end chain length. In the case of methanol and benzene, the distribution was compared with the worm-like chain (WLC)^{41,42} model given by

$$P(R) = \frac{4\pi NR^2}{l_p^2 A^{9/2}} \exp\left(-\frac{3l_c}{4l_p A}\right) \quad (1)$$

where R is the coordinate along the contour of the tail, N is the normalization factor, l_p is the persistence length, l_c is the contour

length and A is given by

$$A = 1 - \frac{R^2}{l_z^2} \quad (2)$$

Translational and Rotational Diffusion. The diffusion coefficient of Chl A was calculated using the Einstein relation:⁴³

$$6Dt = \lim_{t \rightarrow \infty} \langle |\mathbf{r}_i(t) - \mathbf{r}_i(0)|^2 \rangle \quad (3)$$

where $\mathbf{r}_i(t)$ is the coordinate vector of the particle i at time t , and $\mathbf{r}_i(0)$ the coordinate vector of the particle i at time $t = 0$.

Besides the translational diffusion, rotational diffusion provides useful information on how a solute interacts with the solvents. The rotational diffusion of the chlorin ring was calculated using the autocorrelation function of the vector normal to the plane of the ring:

$$C_2(t) = \langle P_2(\mathbf{n}(0) \cdot \mathbf{n}(t)) \rangle \quad (4)$$

where P_2 is the Legendre polynomial of the order 2, \mathbf{n} is the unit vector pointing out of the plane of the ring, and the brackets indicate the average along the trajectory.⁴⁴ The plane of the ring was determined using the atoms that show the least fluctuations in the ring.

Rotational Relaxation Time Constants. The correlation function measured in the experiments is usually approximated by⁴⁵

$$C_2(t) = a \exp[-(6D_r t)] = a \exp(-t/\tau_2) \quad (5)$$

and τ_2 is the rotational relaxation time. Thus, the relaxation time obtained from the simulations is related to the rotational diffusion coefficient by

$$\tau_2 = \frac{1}{6D_r} \quad (6)$$

Viscosity of the Solvents. To compare the rotational relaxation time constants obtained from the simulations with the experimentally determined time constants, the viscosity of the solvents used in the simulations and the experiments has to be taken into account. The viscosities of the solvents used in the simulations were computed from the nonequilibrium MD simulations.⁴⁶

In this method, a sinusoidally varying acceleration, with the profile given by

$$a_x(z) = A \cos(2\pi z/l_z) \quad (7)$$

where A is the amplitude of the acceleration and l_z is the height of the box, was applied in the x direction. In these simulations the length of the boxes in z -direction were set three times longer than in the other directions. The generated velocity profile due to the acceleration can be written as

$$v_x(z) = V \cos(2\pi z/l_z) \quad (8)$$

where V is the amplitude of the generated velocity. The viscosity was then calculated using the relation:

$$\eta = \frac{A}{V} \rho (l_z/2\pi)^2 \quad (9)$$

where ρ is the density of the solvent.

Different simulations were done varying the amplitude of the acceleration. The viscosity at the equilibrium was determined by interpolation.

Solvation Geometry and Energetics. The distribution of solvent molecules around the Mg plays an important role in solvation and solvent-mediated aggregation of Chl A molecules. The pair correlation function, $g_{x,y}(r)$, and the spatial distribution function (SDF)⁴⁷ were used to get insight into the local ordering of the solvent molecules. The subscripts x and y in $g_{x,y}(r)$ denote the particle types, and r denotes the radial distance between the particles x and y . The number of solvent molecules in the different solvation shells of the Mg was calculated using the running integration number (RIN):

$$n = 4\pi\rho_0 \int_0^R g_{Mg,X}(r)r^2 dr \quad (10)$$

where X denotes either O (oxygen atom) or C (carbon atom), and ρ_0 is the number density of the solvent molecule of which the RIN is calculated.

The anisotropic distribution of the solvent atoms around the chlorin ring was analyzed using the SDFs calculated in the Cartesian coordinate system with the origin of the system fixed to the Mg, two of the vectors defined by the vectors joining Mg to the nitrogen atoms and a third vector orthogonal to the plane defined by the first two vectors.

The PMFs of the interaction of the solvent molecules with the Mg, as a function of radial distance, were calculated from the pair correlation functions obtained from the NVT simulations using the relation:³⁸

$$g_{Mg,X}(r) = \exp\left(-\frac{w(r)}{kT}\right) \quad (11)$$

where $w(r)$ is the PMF, k is the Boltzmann constant, and T is the temperature.

Residence Time of Water and Methanol. The lifetime of the contact between the Mg and the solvent molecules in a given solvation shell provides important information about the solvation dynamics. This information can be obtained from MD trajectories in different manners.⁴⁴ A common approach is the use of the so-called survival time correlation function:⁴⁸

$$P_\alpha(t) = \sum_{j=1}^N \sum_{t'} P_{\alpha,j}(t', t' + t) \quad (12)$$

where the probability function, $P_{\alpha,j}(t', t' + t)$ is a binary function that adopts a value of 1 if the solvent molecule labeled j has been in the referred solvation shell around site α from time t' to time $t' + t$, without escaping in this time interval (or leaving the shell during this interval for a time not longer than a t^* interval), and 0 otherwise. The value of $P_{\alpha,j}(t, t + t')$ is averaged over time and over all solvent molecules from conformations sampled from the MD simulation. $P_{\alpha,j}(t=0)$ equals the average number of solvent molecules belonging to the solvation shell of the site j (i.e., the coordination number), and $P_{\alpha,j}(t)$ gives the average number of solvent molecules that still remain in the hydration shell after a time t from when they first entered the shell. The relaxation trend of $P_{\alpha,j}(t)$ provides information about the local dynamics of the solvation molecules. The value of $P_{\alpha,j}(t)$ can be approximated to an exponential function:⁴⁸

$$C_\alpha(t) = A \exp(-t/\tau) \quad (13)$$

where τ is the residence time.

RESULTS AND DISCUSSION

Structural Properties. The cluster analysis of the chlorin ring done with cutoff of 0.02 nm gave 11, 10, and 9 clusters from 5000 sampled structures in water, methanol, and benzene, respectively. The cumulative number of the clusters (Figure 2a) reaches a plateau indicating a good sampling of the conformational space. The first three clusters account for 99% (80, 10, and 9%), 99% (79, 11, and 9%), and 99% (80, 14, and 5%) of the structures in the three solvents, respectively. The representative structures of the first three clusters are shown in Figure 3. The average structure of the ring is planar. The deviations from the planarity involve different collective motions of the atoms in the ring. The root-mean-square fluctuations of the atoms (Figure 4) show that

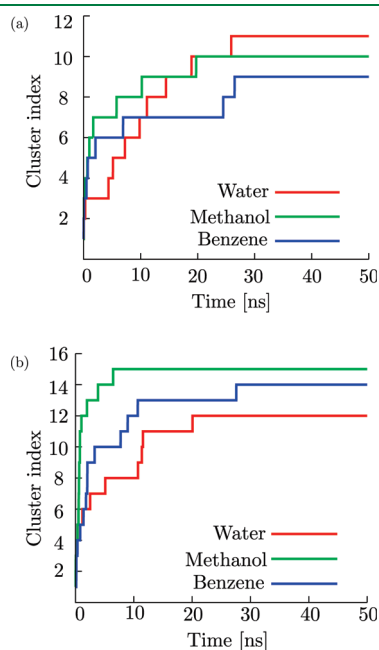


Figure 2. Cumulative distribution of the number of clusters obtained from the water, methanol, and benzene simulations: (a) chlorin ring and (b) all heavy atoms of Chl A.

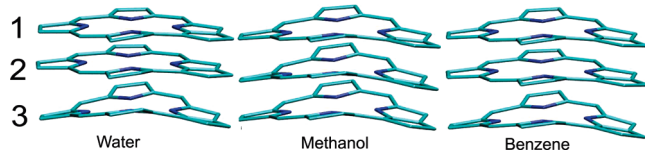


Figure 3. Representative configurations of the three most populated clusters of the chlorin ring.

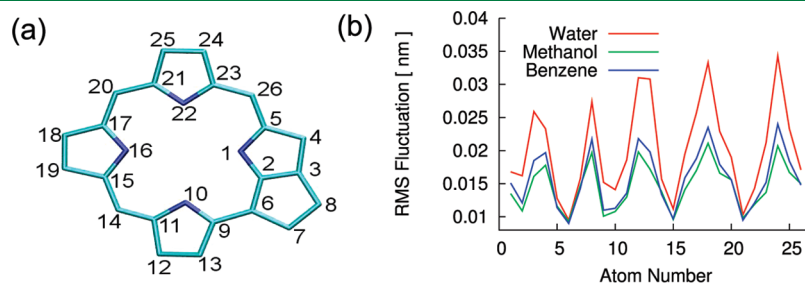


Figure 4. Root-mean-square fluctuations of the atoms of the chlorin ring (b). The numbering of the atoms is shown in (a).

atoms on the border of the chlorin ring have larger fluctuations than the atoms that connect the aromatic rings or the atoms in the inner part of the rings. Though the pattern of fluctuations in all the three solvents is similar, larger fluctuations are observed in water than in methanol or benzene due to the phytol tail which folds back onto the chlorin ring.

The cluster analysis of the complete molecule performed using a larger cutoff of 0.3 nm gave 12, 15, and 14 clusters from 5000 structures sampled from the simulations in water, methanol, and benzene, respectively. The cumulative number of the clusters (Figure 2b) reaches a plateau indicating a good sampling of the conformational space. The first three clusters account for 94, 71, and 75% of the total sampled structures in water, methanol, and benzene, respectively. The large variety of conformations is mainly determined by the hydrophobic phytol chain. The smaller number of clusters observed in the simulation in water is due to the folded configuration of the tail, which reduces its mobility. On the contrary, in methanol and benzene, the flexibility of the tail resembles that of a freely floating chain thereby increasing the number of conformations. In Figure 5 the beginning-to-end distribution of the phytol chain is reported. In methanol and benzene, the distributions are similar, spanning from 0.36 to 2.33 nm, with the main peak in methanol at 1.57 nm and in benzene at 1.76 nm. The WLC (eq 1) model fitted to the distributions gives contour and persistence lengths of $l_c = 2.47 \pm 0.02$ nm, $l_p = 0.239 \pm 0.006$ nm, and $l_c = 2.575 \pm 0.009$ nm, $l_p = 0.266 \pm 0.003$ nm in methanol and benzene, respectively. The comparison of the persistence lengths indicates that the chain in benzene is stiffer than in methanol. The distribution in water, which does not fit to the WLC model, is bimodal with the first peak at 0.57 nm corresponding to the tail folded onto the ring and the second peak at 1.4 nm corresponding to a more extended configuration.

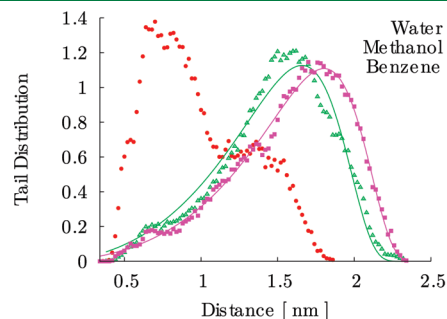


Figure 5. The distribution of the beginning-to-end distance of the phytol chain of Chl A in water, methanol, and benzene. The symbols indicate the data obtained from the simulations. The distribution in benzene and methanol is fitted to the worm-like chain model distribution (solid lines).

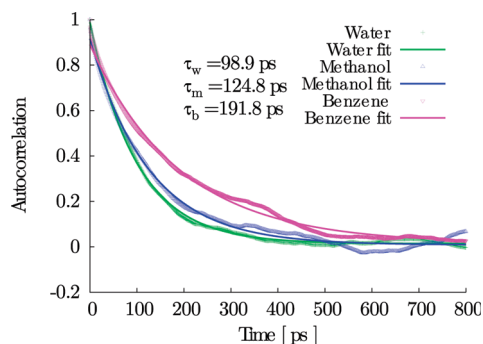


Figure 6. Correlation functions using second-order Legendre polynomial for the rotational diffusion of Chl A in water, methanol, and benzene. The correlation functions are fitted with exponential decay curves. The time constants for the decay, τ , are reported in the figure.

Linear and Rotational Diffusion. The rotational diffusion of a molecule in solution can be measured in real time using pump–probe techniques with ultrashort laser pulses. In these experiments, a pump pulse is used to select molecules with a defined orientation, in some cases the molecules can also be aligned with the laser pulses, and the probe pulse is used to interrogate the transient anisotropy in the system induced by the pump pulse. Rotational diffusion makes the system isotropic thereby diminishing the signal from the induced anisotropy. Few of these experiments have been done to measure the rotational dynamics of chlorophyll molecules. In one of the transient grating studies done on zinc methyl 13-desoxyphytylphorbide (Zn DMPPH), an analogous molecule, the measured relaxation time of the rotation of the molecule dissolved in tetrahydrofuran (THF) was 114 ps.⁴⁹ The relaxation time of the rotational dynamics of a molecule is directly proportional to the viscosity, η , of the solvent. Thus, to compare the experimental result with the results from simulations, the viscosities of the different solvents need to be taken into account. The experimental viscosity of THF is $\eta_T = 4.8 \times 10^{-4}$ Pa.s. The viscosities of the solvent models used in the simulations, determined by nonequilibrium MD simulations,⁵⁰ are $\eta_w = (5.6 \pm 0.7) \times 10^{-4}$ Pa.s, $\eta_m = (5.4 \pm 0.7) \times 10^{-4}$ Pa.s, and $\eta_b = (7.4 \pm 0.9) \times 10^{-4}$ Pa.s for water, methanol, and benzene, respectively.

In the approximation that the geometry of the molecule is unchanged in different solvents, the viscosities and the rotational time constants τ follow the relation:

$$\frac{\tau_x}{\tau_y} = \frac{\eta_x}{\eta_y} \quad (14)$$

where subscripts x and y represent two different solvents. The expected rotational time constant of Zn DMPPH in water, methanol, and benzene models estimated using eq 14 are $\tau_{2,w} = 135 \pm 15$, $\tau_{2,m} = 128 \pm 14$, and $\tau_{2,b} = 176 \pm 30$ ps, respectively. Subscripts w, m, and b refer to the solvents water, methanol, and benzene, respectively. The time constants of the exponential functions used to fit the second-order correlation functions obtained from the simulations (Figure 6) are $\tau_w = 99$, $\tau_m = 125$, and $\tau_b = 192$ ps. The rotational diffusion time constants obtained from simulations in methanol and benzene are close to the respective estimated rotational time constants. In water, the time constant obtained from the simulation is slightly lower than the corresponding estimate from the experiment, which could be due to the influence of the solvent in the geometry of the

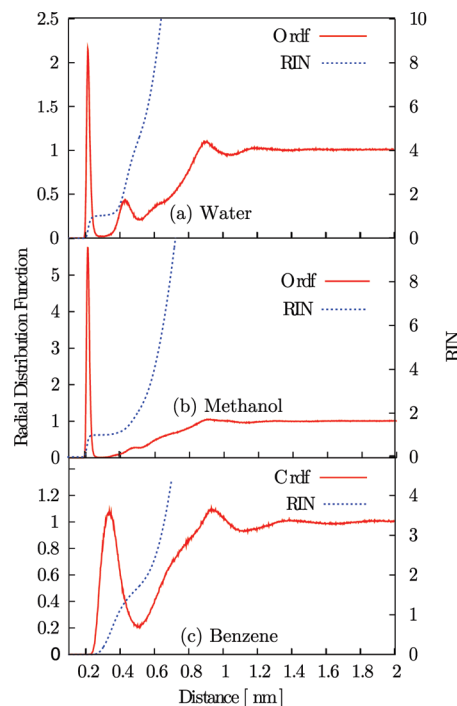


Figure 7. The radial distribution functions and the running integration numbers of the oxygen atoms of water and methanol and the carbon atoms of benzene around the Mg atom.

molecule. In water, as noted previously, the tail folds back to the chlorin ring, thereby decreasing the cross-section of the molecule as well as the moment of inertia and hence the faster rotational motion.

The rotational diffusion coefficients computed using eq 6 are 1.68, 1.33, and 0.87 rad^2/ns in water, methanol, and benzene, respectively. To the best of our knowledge, direct experimental determination of rotational diffusion coefficients of Chl A or related molecules is not yet done. However, the values obtained from the simulations can be considered quite reliable, as the relaxation times obtained from the simulations are close to those estimated from the available experimental data.

Finally, the linear diffusion constants estimated using eq 3 are 0.47×10^{-5} , 0.52×10^{-5} , and 0.40×10^{-5} cm^2s^{-1} in water, methanol, and benzene, respectively.

Chl A Solvation. The coordination unsaturation of the Mg can be satisfied by nucleophilic ligands.² The ligands can form both the $\text{Chl} \cdot \text{L}_1$ and $\text{Chl} \cdot \text{L}_2$ complexes, where $\text{Chl} \cdot \text{L}_1$ is the complex formed with a ligand occupying one of the axial positions, while $\text{Chl} \cdot \text{L}_2$ is the complex formed with two ligands occupying the axial positions on both the sides of the chlorin plane. The coordination of the methanol and water molecules to the Mg can be studied with the pair correlation function $g_{\text{Mg},\text{O}}$. In Figure 7, the $g_{\text{Mg},\text{O}}$ s of both the solvents are shown. They have the first peak located at 0.21 nm. The narrow width of the peaks indicates that the oxygen atom is strongly bound to the Mg in both the solvents. The average number of oxygen atoms within the first solvation shell (up to 0.28 nm) is 1.02 and 1.01 for water and methanol, respectively (see Figure 7). This indicates that the Mg is predominantly penta-coordinated. The average number of oxygen atoms within the second solvation shell (up to 0.50 nm in water and 0.52 nm in methanol) is 4.33 and 2.19 in water and methanol, respectively.

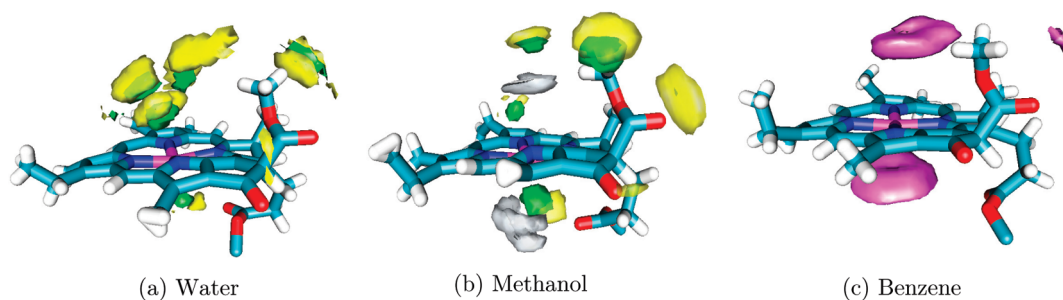


Figure 8. Spatial distribution function of oxygen and hydrogen atoms of water molecules (a), methanol molecules (b), and carbon atoms of benzene molecules (c) around the chlorin ring. Color coding for the contour surfaces: green for O, yellow for H in hydroxyl group, gray for H in methyl group, and magenta for C in the benzene. The contour values of the iso surfaces are 10 for both the hydrogens and oxygens in water; 22, 9, and 15 for methyl hydrogens, hydroxyl hydrogens, and oxygens, respectively, in methanol; and 12 for the carbons in benzene.

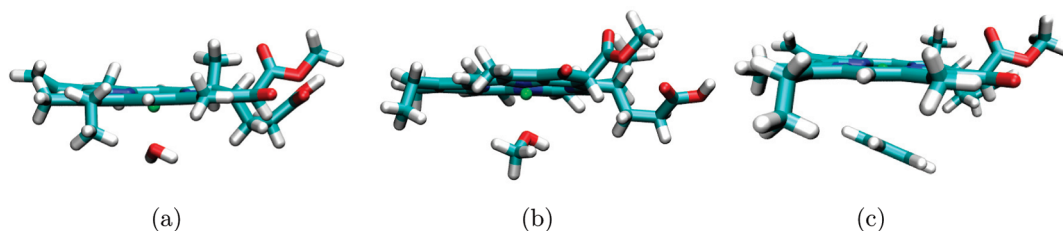


Figure 9. The optimized geometry of chlorophyllide–water (a), chlorophyllide–methanol (b), and chlorophyllide–benzene (c) complexes. The oxygen atom of water and methanol is bound to the Mg atom of the chlorin ring. The distance between the oxygen and the Mg is 0.21 nm (a and b). The distance of the Mg from the center of the benzene ring is 0.38 nm. The plane of the benzene ring is slightly tilted at an angle to the plane of the chlorin ring (c).

In Figure 8a and b, the SDF of oxygen and hydrogen atoms of water and methanol around the chlorin ring is shown. The average structures of the chlorin ring shown in the figures are slightly distorted with the Mg atom displaced out of the plane of the ring. The distortion is opposite to the side with the ester C=O group. The distances between the Mg and the plane of the chlorin ring defined by atoms 1, 10, and 22 (Figure 4a) are 0.023 and 0.027 nm in water and methanol, respectively. The B3LYP/6-31G** optimized structures of water–chlorophyllide and methanol–chlorophyllide complexes (Figure 9a and b, respectively) also show distortion of the chlorin ring. The distances between the Mg atom and the plane of the chlorin ring, as defined previously, are 0.040 and 0.037 nm in the water–chlorophyllide and the methanol–chlorophyllide complexes, respectively. The displacement observed in the optimized geometry of water–chlorophyllide complex is similar to the displacement of 0.039 nm observed in the crystal structure of ethyl–chlorophyllide dihydrate.¹⁹ The displacement observed in the MD simulations is less than the displacement observed in the chlorophyllide crystal structure and close to the range of displacement of 0.011–0.025 nm observed in the chlorophyll molecules of the spinach major LHC crystal structure (1RWT).

The SDF of the oxygen atoms of water (Figure 8a) shows two densities on the axial position on both the sides of the chlorin ring. In the same figure, the SDF of the hydrogen atoms of water shows that the hydrogens of the water molecules bound to the Mg point outward from the ring, which is consistent with the B3LYP/6-31G** optimized structure of water–chlorophyllide complex as shown in Figure 9a. The distance between Mg and oxygen in the optimized structure is 0.211 nm, which agrees with the position of the first peak of the $g_{\text{Mg,O}}$ from the simulation. In addition, the Mg–O distances in both the QM-optimized geometry and the SDF are similar to the distance of 0.204 nm observed in the crystal structure of ethyl chlorophyllide a

dihydrate.¹⁹ The SDF of the oxygen atoms shows other two high-density regions above the first on the upper side of the ring. The arrangement of these two regions indicates presence of hydrogen bonds between the water molecules in the two regions and the water molecules coordinated to the Mg. Aggregation of chlorophylls in aqueous medium has been attributed to the hydrogen bonding between the Mg-bound water molecule of one chlorophyll and the keto C=O group of another chlorophyll molecule.^{51,52} The presence of water molecules chained with hydrogen bonds to the Mg coordinated water molecule is supported by the crystal structure of ethyl chlorophyllide a · 2H₂O, as proposed by Strouse et al.¹⁹ In this crystal structure, the Mg-coordinated water molecule is simultaneously hydrogen bonded to the keto C=O group of another ethyl chlorophyllide a and to the oxygen of the second water molecule. The second water molecule is then hydrogen bonded to the ester C=O of the first ethyl chlorophyllide and to the propionic ester C=O of a third ethyl chlorophyllide (see the cited paper of Strouse et al.).¹⁹ Coordination of water molecule to the Mg also plays a central role in the proposed models of the photoactive chlorophyll special pair;⁵² the SDF of the water molecules around the chlorin ring obtained from our simulations is consistent with these hypotheses.

The SDF of the oxygen of methanol (Figure 8b) shows three regions of high density of oxygen atoms: two above and one below the chlorin ring. The high densities on the axis above and below the Mg are due to the oxygen atoms bound to the Mg; the other high density on the side of the density above the Mg is due to methanol molecules hydrogen bonded to the former. The SDF of the hydroxyl hydrogens shows a high density in the vicinity of the ester and the keto oxygens indicating hydrogen bonding of methanol molecules with these groups. How the arrangement of the two methanol molecules affects the solubility/aggregation of chlorophylls is not clear from the simulations.

The pair correlation function $g_{\text{Mg,C}}$ of carbon atoms of benzene molecules around the Mg has the first peak at 0.35 nm (Figure 7c), which is similar to the distance of 0.38 nm between Mg and benzene in the B3LYP/6-31G** optimized chlorophyllide–benzene complex (Figure 9c). The number of solvent molecules in the first solvation shell (up to 0.5 nm radial distance) is 1.65. As in the case of water and methanol, the Mg is slightly displaced out of the plane of the ring by about 0.013 nm. The displacement of the Mg observed in the B3LYP/6-31G** optimized chlorophyllide–benzene (Figure 9c) complex is 0.006 nm. The oblate doughnut shape densities, which slightly are at an angle to the chlorin plane in the SDF of benzene (see Figure 8c), agree well with the position of the benzene molecule in the QM-

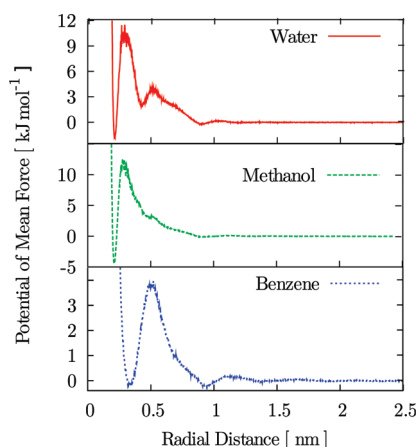


Figure 10. The potential of the mean force of the solvent molecules derived from the radial distribution functions.

optimized structure of chlorophyllide–benzene complex (see Figure 9c).

Binding Energy and Helmholtz Free Energies of Solvation of the Mg Atom. The binding energies of water and methanol molecules in the chlorophyllide $A \cdot L_1$ calculated using counterpoise corrected B3LYP/6-31G** method are -43.73 and -45.34 kJ/mol, respectively. QM calculations done by Fredj et al.¹⁸ on a similar model of Chl A gave a similar value of -51.46 kJ/mol, for the binding of water to the Mg. The Helmholtz free energy of binding (ΔF_b) obtained from our calculations is -7.95 and -7.96 kJ/mol in water and methanol, respectively. From the experimentally determined equilibration constant ($K = 56$ l mol⁻¹) of dimers and the methanol-coordinated Chl A molecules in CCl_4 ,¹⁷ a ΔF_b value of -10.04 kJ/mol can be estimated. As we expect that the ΔF_b for the binding of methanol to chlorophyllide a is similar to the one with Chl A, the QM calculated value is in good agreement with the experimental data.

The ΔF_b calculated from the MD simulations using the PMF (Figure 10) of the interaction between the methanol molecules and the Mg is -4.45 kJ/mol, which is lower than the value estimated from the experiment. However, it has been shown from QM calculations that the dielectric constant of the medium decreases the binding of the ligands.¹⁸ Thus taking into account the difference in the dielectric constants of the solvents in the experiment (CCl_4 , $\epsilon = 2.24$) and the simulations (pure methanol, $\epsilon \approx 20$ ⁵³), the lower values of the ΔF_b in the simulation can be considered reasonable.

The ΔF_b of water and benzene to the Mg obtained from MD simulations is -1.92 and -0.24 kJ/mol. Though experimental values are not available for these solvents, the values obtained from our simulation seem reasonable.

Chl A Aggregation. The structures of dimers and multimers obtained from the simulation of 10 Chl A molecules in the water

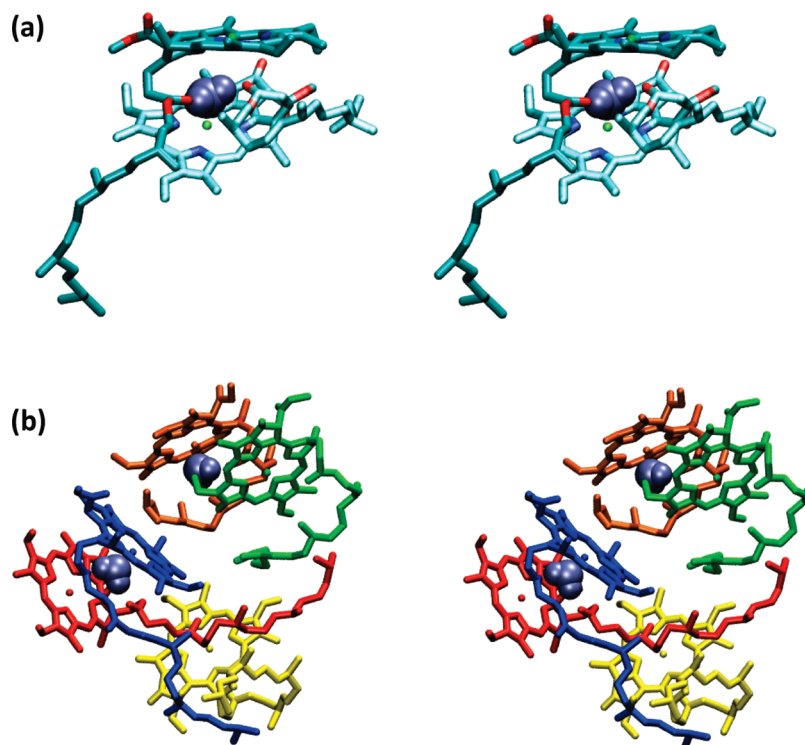


Figure 11. Stereo view of the structures of the dimer (a) and the pentamer (b) of Chl A formed in the water.

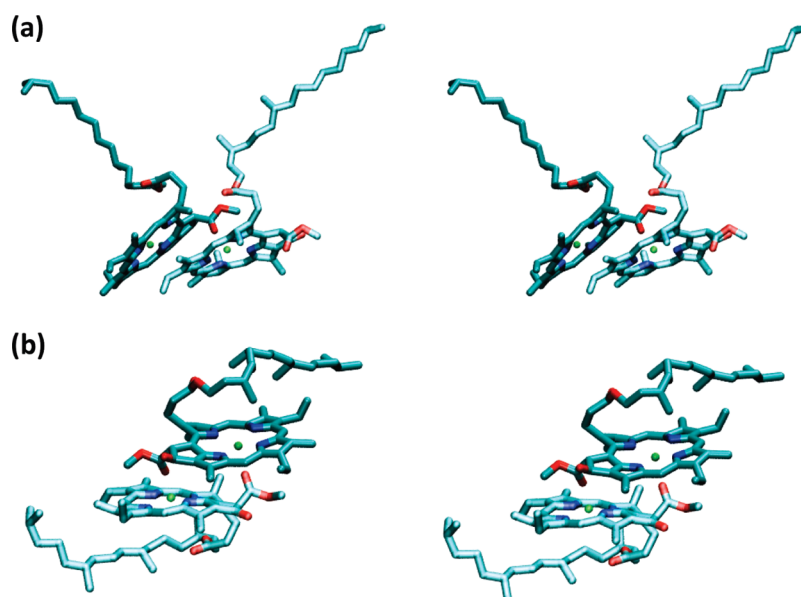


Figure 12. Stereo view of the structures of the dimers of Chl A in the benzene.

and the benzene are shown in the Figures 11 and 12, respectively. No stable dimers or multimers were observed in the simulation in methanol. In the simulation in water, a stable dimer and a pentamer were observed. The spherical shape of the pentamer with the chlorin rings pointing outward is similar to the structure of the large aggregates of Chl A in water-rich regions, as proposed by Agostiano et al.,⁵ wherein the self-aggregation is attributed to the hydrophobic interactions of the phytol tail. The dimer obtained from the water simulation (Figure 11a) has a water molecule sandwiched between the two Chl A molecules, while the pentamer in water (Figure 11b) has two water molecules sandwiched between the Chl A molecules. It is possible that the water molecules trapped in the aggregates also contribute to the aggregation by electrostatic and hydrogen-bonding interactions.

The dimers observed in the simulation in benzene (Figure 12) have different geometries. In the first dimer (Figure 12a), the two molecules are bridged by the binding of the ester C=O group of the first molecule to the Mg of the second molecule. In the second dimer (Figure 12b), the two molecules are bridged by the binding of the ester C=O group of one molecule to the other. Though other dimeric and multimeric conformations are also expected to form,⁵⁴ they probably were not observed in our simulations because of the limited size of the system and the length of the simulations.

Residence Time of Water and Methanol. As reported by Ballschmimer and Katz,⁵¹ a dynamic equilibrium exists between the 2Chl A·L complex and the isolated species for nucleophilic ligand L in nonpolar solvents. The coordination of the ligand to the Mg is primarily responsible for the formation of Chl A·L complex. The time scales of the formation of the complex and its disaggregation can be calculated from the residence time of methanol and water coordinated to the Mg.

In the simulation in methanol, the solvent molecules in close contact to the Mg have residence times spanning from picoseconds up to few nanoseconds. The distribution of the short residence time is shown in Figure 1a of the Supporting Information. The exponential decay approximating the distribution gives a time constant of 5 ps. The distribution of the residence time longer than 100 ps is shown in the panel b of the same figure. The

exponential function fitted to the distribution gives a time constant of 2566 ps.

A similar multi-exponential distribution is observed for the residence times of water molecule in the chlorin ring (see Figure 2 in the Supporting Information). Interestingly, in this case, intermediate time scales (hundreds of ps) are also observed. The three time scales obtained from fitting exponential functions are 10, 200, and 1300 ps, respectively. The shortest and the longest time scales are associated with events involving a fast solvent exchange in the first shell of the Mg caused by molecular collision or conformational change of the chlorin ring. Interaction of other water molecules hydrogen bonded to the Mg-coordinated molecule may be responsible for the presence of the intermediate residence time. However, as there are few data points available for the distributions of the residence times, the calculated decay constants can only be considered qualitatively.

CONCLUSIONS

In this paper, we have presented a new model of chlorophyll a for MD simulations based on the all-atom OPLS force field. The model was tested by studying structural and dynamic properties of the molecule in three different solvents: water, methanol, and benzene. The rotational time constants obtained with our model from the simulations in methanol (125 ps) and benzene (192 ps) are in good agreement with the value extrapolated from experimental data.

The distribution of the phytol tail length in methanol and in benzene is consistent with WLC distribution. The stark differences in the configuration of the tail in different solvents raise interesting questions about the effect of the protein and lipid environment on the distribution of the Chl A in the photosynthetic complex of the chloroplasts. Furthermore, the Mg-coordinating water and methanol molecules have geometries comparable to the available experimental values. In particular, the spatial distribution of the oxygens and hydrogens of water and the methanol molecules support the proposed interaction of these solvent molecules with different functional groups of Chl A. The results of the simulations show that structural and dynamic

properties of the chlorin ring are similar in both methanol and benzene. In methanol and water, the Mg atom in the chlorin ring binds the oxygen of the solvent molecules with residence times of 2566 and 1300 ps, respectively.

The overall good quality of the model makes it suitable for the study of interesting systems like self-assembled micelles,⁵⁵ mimicking the reaction center of light harvesting complexes and interactions of the chlorophylls with inorganic surfaces.

■ ASSOCIATED CONTENT

S Supporting Information. The coordinates of the QM optimized structures of Chl A, chlorophyllide A, complexes of chlorophyllide A with the solvent molecules, the bonded and nonbonded parameters used in the simulations, and the histograms of the residence time of the solvent molecules in the chlorin ring. This information is available free of charge via the Internet at <http://pubs.acs.org/>

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: d.rocacatano@jacobs-university.de.

■ ACKNOWLEDGMENT

This study was performed using the computational resources of the Computer Laboratories for Animation, Modeling and Visualization (CLAMV) at Jacobs University Bremen. This work was performed within the graduate program “Nanomolecular Science” and was financially supported by “Research Center for Functional Materials and Nanomolecular Science (NANOFUN)”.

■ REFERENCES

- (1) Katz, J. J.; Oettmeier, W.; Norris, J. R. *Phil. Trans. R. Soc. Lond. B* **1976**, *273*, 227–253.
- (2) Katz, J. J. *Naturwissenschaften* **1973**, *60*, 32–39.
- (3) Agostiano, A.; Cosma, P.; Trotta, M.; Monsu-Scolaro, L.; Micali, N. *J. Phys. Chem. B* **2002**, *106*, 12820–12829.
- (4) Emerson, R.; Arnold, W. J. *Gen. Physiol.* **1931–1932**, *16*, 191–205.
- (5) Agostiano, A.; Monica, M. D.; Palazzo, G.; Trotta, M. *Biophys. Chem.* **1993**, *47*, 193–202.
- (6) Jiao, J.; Thamyongkit, P.; Schmidt, I.; Lindsey, J. S.; Bocian, D. F. *J. Phys. Chem. C* **2007**, *111*, 12693–12704.
- (7) Anariba, F.; Viswanathan, U.; Bocian, D. F.; McCreery, R. L. *Anal. Chem.* **2006**, *78*, 3104–3112.
- (8) Thamyongkit, P.; Yu, L.; Padmaja, K.; Jiao, J.; Bocian, D. F.; Lindsey, J. S. *J. Org. Chem.* **2006**, *71*, 1156–1171.
- (9) Jiao, J.; Anariba, F.; Tiznado, H.; Schmidt, I.; Lindsey, J. S.; Zaera, F.; Bocian, D. F. *J. Am. Chem. Soc.* **2006**, *128*, 6965–6974.
- (10) Zannoni, R.; Aurora, A.; Cattaruzza, F.; Decker, F.; Fastiggi, P.; Menichetti, V.; Tagliatesta, P.; Capodilupo, A.-L.; Lembo, A. *Mater. Sci. Eng., C* **2007**, *27*, 1351–1354.
- (11) Simkiene, I.; Sabataityte, J.; Babonas, G. J.; Reza, A.; Beinoras, J. *Mater. Sci. Eng., C* **2006**, *26*, 1007–1011.
- (12) Linnanto, J.; Korppi-Tommola, J. *Phys. Chem. Chem. Phys.* **2006**, *8*, 663–687.
- (13) Damjanovic, A.; Kosztin, I.; Kleinekathofer, U.; Schulten, K. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **2002**, *65*, 1–24.
- (14) Olbrich, C.; Kleinekathofer, U. *J. Phys. Chem. B* **2010**, *114*, 12427–12437.
- (15) Spezia, R.; Aschiy, M.; Nola, A. D.; Valentin, M. D.; Carbonera, D.; Amadei, A. *Biophys. J.* **2003**, *84*, 2805–2813.
- (16) Palencar, P.; Vacha, F.; Kutz, M. *Photosynthetica* **2005**, *43*, 417–420.
- (17) Katz, J. J.; Strain, H. H.; Leussing, D. L.; Dougherty, R. C. *J. Am. Chem. Soc.* **1968**, *90*, 784–791.
- (18) Fredj, A. B.; Ruiz-Lopez, M. F. *J. Phys. Chem. B* **2010**, *114*, 681–687.
- (19) Chow, H.-C.; Serlin, R.; Strouse, C. E. *J. Am. Chem. Soc.* **1975**, *97*, 7230–7237.
- (20) Canc_es, E.; Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3032–41.
- (21) Miertu_s, S.; Scrocco, E.; Tomasi, J. *J. Chem. Phys.* **1981**, *55*, 117–129.
- (22) Liu, Z.; Yan, H.; Wang, K.; Kuang, T.; Zhang, J.; Gui, L.; An, X.; Chang, W. *Nature* **2004**, *428*, 287–292.
- (23) Breneman, C. M.; Wiberg, K. B. *J. Comput. Chem.* **1990**, *11*, 361–397.
- (24) 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, J. B.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, revision A.1; Gaussian Inc.: Wallingford, CT, 2009.
- (25) Jorgensen, W. L.; Maxwell, D. S.; Tirald-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.
- (26) Jorgensen, W. L. Chapter OPLS, force field. *Encyclopedia of Computational Chemistry*; Wiley: New York, 1998; Vol. 3, pp 1986–1989.
- (27) Jorgensen, W. L.; McDonald, N. A. *J. Mol. Struct. (THEOCHEM)* **1998**, *424*, 145–155.
- (28) Jorgensen, W. L.; McDonald, N. A. *J. Phys. Chem. B* **1998**, *102*, 8049–8059.
- (29) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (30) Milano, G.; Mller-Plathe, F. *J. Phys. Chem. B* **2004**, *108*, 7415–7423.
- (31) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.
- (32) Hoover, W. G. *Phys. Rev. A: At., Mol., Opt. Phys.* **1985**, *31*, 1695–1697.
- (33) Parrinello, M.; Rahman, A. *J. Apply. Phys.* **1981**, *52*, 7182–7190.
- (34) Nos_e, S.; Klein, M. L. *Mol. Phys.* **1983**, *50*, 1055–1076.
- (35) Miyamoto, S.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 952–962.
- (36) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J. Comput. Chem.* **1997**, *18*, 1463–1472.
- (37) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8592.
- (38) Chandler, D. In *Introduction to modern statistical mechanics*; Oxford University Press: Berkeley, CA, 1987; Chapter 7, page 201.
- (39) Smith, L. J.; Daura, X.; van Gunsteren, W. *Proteins: Struct., Funct., Genet.* **2002**, *48*, 487–496.
- (40) Daura, X.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. *J. Mol. Biol.* **1998**, *280*, 925–932.
- (41) Thirumalai, D.; Ha, B. *Statistical Mechanics of Semixible Chains: A Meanfield Variational Approach. Theoretical and Mathematical Models in Polymer Research*; Academic Press: San Diego, CA, 1998; pp 1–35.

- (42) Becker, N. B.; Rosa, A.; Everaers, R. *Eur. Phys. J. E: Soft Matter Biol. Phys.* **2010**, *32*, 53–69.
- (43) Allen, M. P.; Tildesley, D. J. *Statistical Mechanics. Computer Simulations of Liquids*; Oxford Science Publications: Oxford, England, 1987; pp 58–60.
- (44) Roccatano, D. *Curr. Protein Pept. Sci.* **2008**, *9*, 407–426.
- (45) Jas, G. S.; Wang, Y.; Pauls, S. W.; Johnson, C. K.; Kuczera, K. *J. Chem. Phys.* **1997**, *107*, 8800–8812.
- (46) Hess, B. *J. Chem. Phys.* **2002**, *116*, 209–217.
- (47) Kulasik, P. G.; Laaksonen, A.; Svishchev, I. M. *Spatial Structure in Molecular Liquids. Molecular dynamics: from classical to quantum methods*; Elsevier Science B.V: Amsterdam, The Netherlands, 1999; pp 61–90.
- (48) Impey, R. W.; Madden, P. A.; McDonald, I. R. *J. Phys. Chem.* **1983**, *87*, 5071–5083.
- (49) Wiederrecht, G. P.; Svec, W. A.; Niemczyk, M. P.; Wasielewski, M. R. *J. Phys. Chem.* **1995**, *99*, 8918–8926.
- (50) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. *J. Comput. Chem.* **2005**, *26*, 1701–1718.
- (51) Ballschmiter, K.; Katz, J. J. *J. Am. Chem. Soc.* **1969**, *91*, 2661–2677.
- (52) Katz, J. J.; Norris, J. R.; Shipman, L. L.; Thurnauer, M. C. *Annu. Rev. Biophys. Bioeng.* **1978**, *7*, 393–434.
- (53) Richardi, J.; Millot, C.; Fries, P. H. *J. Chem. Phys.* **1999**, *110*, 1138–1147.
- (54) Oba, T.; Furukawa, H.; Wang, Z.-Y.; Nozawa, T.; Mimuro, M.; Tamiaki, H.; Watanabe, T. *J. Phys. Chem. B* **1998**, *102*, 7882–7889.
- (55) Worcester, D. L.; Michalski, T. J.; Katz, J. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 3791–3795.

Polarizable Atomic Multipole X-Ray Refinement: Particle Mesh Ewald Electrostatics for Macromolecular Crystals

Michael J. Schnieders,^{*,†} Timothy D. Fenn,^{‡,§} and Vijay S. Pande[†]

[†]Department of Chemistry

[‡]Department of Molecular and Cellular Physiology

Stanford University, Stanford California 94305, United States

[§]Howard Hughes Medical Institute, Chevy Chase, MD 20815-6789, United States

S Supporting Information

ABSTRACT: Refinement of macromolecular models from X-ray crystallography experiments benefits from prior chemical knowledge at all resolutions. As the quality of the prior chemical knowledge from quantum or classical molecular physics improves, in principle so will resulting structural models. Due to limitations in computer performance and electrostatic algorithms, commonly used macromolecules X-ray crystallography refinement protocols have had limited support for rigorous molecular physics in the past. For example, electrostatics is often neglected in favor of nonbonded interactions based on a purely repulsive van der Waals potential. In this work we present advanced algorithms for desktop workstations that open the door to X-ray refinement of even the most challenging macromolecular data sets using state-of-the-art classical molecular physics. First we describe theory for particle mesh Ewald (PME) summation that consistently handles the symmetry of all 230 space groups, replicates of the unit cell such that the minimum image convention can be used with a real space cutoff of any size and the combination of space group symmetry with replicates. An implementation of symmetry accelerated PME for the polarizable atomic multipole optimized energetics for biomolecular applications (AMOEBA) force field is presented. Relative to a single CPU core performing calculations on a P1 unit cell, our AMOEBA engine called Force Field X (FFX) accelerates energy evaluations by more than a factor of 24 on an 8-core workstation with a Tesla GPU coprocessor for 30 structures that contain 240 000 atoms on average in the unit cell. The benefit of AMOEBA electrostatics evaluated with PME for macromolecular X-ray crystallography refinement is demonstrated via rerefinement of 10 crystallographic data sets that range in resolution from 1.7 to 4.5 Å. Beginning from structures obtained by local optimization without electrostatics, further optimization using AMOEBA with PME electrostatics improved agreement of the model with the data (R_{free} was lowered by 0.5%), improved geometric features such as favorable (ϕ , ψ) backbone conformations, and lowered the average potential energy per residue by over 10 kcal/mol. Furthermore, the MolProbity structure validation tool indicates that the geometry of these rerefined structures is consistent with X-ray crystallographic data collected up to 2.2 Å, which is 0.9 Å better than the actual mean quality (3.1 Å). We conclude that polarizable AMOEBA-assisted X-ray refinement offers advantages to methods that neglect electrostatics and is now efficient enough for routine use.

I. INTRODUCTION

We recently described theory for biomolecular X-ray crystallography refinement based on optimization of a target function E_{target} that is the sum of polarizable atomic multipole descriptions of both the X-ray scattering $E_{\text{X-ray}}$ and the chemical potential energy $E_{\text{chemistry}}$

$$E_{\text{target}} = w_a E_{\text{X-ray}} + E_{\text{chemistry}} \quad (1)$$

where the weight w_a controls their relative importance.¹ The method has been successfully applied to ultrahigh-resolution (0.5 Å) peptide crystals¹ and high-resolution (1.0 Å) protein and nucleic acid biomolecules² and to neutron crystallography.³ In the first two cases, our Cartesian Gaussian multipolar scattering model with multipole coefficients from the atomic multipole optimized energetics for biomolecular applications (AMOEBA) force field^{4–8} improved R/R_{free} statistics.^{1,2} For the biomolecular and neutron crystallography data sets, the polarizable AMOEBA energetic model was shown to be critical for refinement of water

hydrogen-bonding networks.^{2,3} These encouraging results motivate further work to apply AMOEBA-assisted X-ray refinement to larger macromolecular crystals at lower resolution (1.0–4.5 Å). However, our rough draft implementation based on the combination of TINKER v. 5.0⁹ and CNS v. 1.21¹⁰ required expansion to P1 for AMOEBA forces and was limited to systems with on the order of 25 000 atoms. To address this, the present work describes a completely new implementation of AMOEBA ($E_{\text{chemistry}}$) designed from the ground up for application to large biomolecular data sets on modern desktop workstations. We describe particle mesh Ewald (PME) electrostatics theory that consistently supports all 230 space groups, replicates of the unit cell such that the minimum image convention can be used with a real space cutoff of any size and the combination of space group symmetry with replicates.¹¹ Our implementation, called Force Field X (FFX), uses the Java Runtime Environment (JRE) for

Received: September 4, 2010

Published: March 09, 2011

shared memory parallelization across CPU cores in combination with offloading computational work to a GPU coprocessor.

A rigorous solution for the electrostatic potential within an infinite lattice of unit cells was originally described by Ewald in 1921.¹² The method, now referred to as Ewald summation, converts the real space Coulomb lattice summation into the sum of real space and reciprocal space contributions. PME, introduced by Darden et al. in 1993,¹¹ formulated the reciprocal space portion using Lagrange interpolation and fast Fourier transforms (FFT) to achieve $N \cdot \log(N)$ scaling for the calculation of structure factors. Smooth PME¹³ replaced Lagrange interpolation with B-spline interpolation, which offers analytic gradients of arbitrary accuracy and extension to multipolar charge descriptions.^{14–16} Recently, Gaussian split Ewald,¹⁷ multilevel Ewald,¹⁸ and interlaced^{19,20} approaches have been presented to improve the scaling and/or parallelization of particle mesh methods.

Computation of structure factors is of fundamental importance to both PME electrostatics and X-ray scattering. In both cases this is a direct consequence of the tiling of three-dimensional (3D) space by a repeating unit cell. In the context of X-ray refinement, structure factors computed from the structural model are formally compared to those measured experimentally within the X-ray term ($E_{X\text{-ray}}$) of the overall refinement target given in eq 1. Earlier work based on the FFX platform presented a differentiable X-ray term that includes the scattering of bulk solvent, which is used for the AMOEBA-assisted rerefine-ments presented later.²¹ Within the context of force field potentials ($E_{\text{chemistry}}$), periodic boundary conditions (PBC) facilitate the study of an infinitely large system and eliminate edge effects inherent to aperiodic descriptions.

Use of PME for molecular dynamics simulations has flourished because it is often the most efficient way to avoid artifacts introduced by truncation schemes.^{14,22} The importance of electrostatics to biomolecular energetics has led to the development of schemes to decompose electron density^{23,24} into relatively many low-order sites (i.e., charges at atomic centers, bond centers, lone pairs, etc.) vs fewer higher order atomic multipole sites.^{25–29} Generally, there has been greater emphasis placed on modeling lone pair electron density than on bonding electron density in order to predict hydrogen bonding.⁵ From a crystallography perspective, bond density is of greater consequence since it contributes more to X-ray scattering than lone pair density for the majority of biomolecular structures.³⁰ For example, the phenol ring of tyrosine has seven bonds between heavy atoms but only a single lone pair site. To place electron density at bond centers for tetrahedral chemistries, diamond for example, either an atomic multipole expansion through hexadecapole order or bond charges is necessary.^{1,31}

On the other hand, use of Ewald summation for X-ray crystallography refinement has lagged behind its adoption for molecular simulation. A first step toward including electrostatics within refinement by simulated annealing was described by Weis et al. for influenza virus hemagglutinin; however, the lattice summation was evaluated using a conditionally convergent spherical cutoff.³² This approach was reintroduced³³ with the addition of an analytic generalized Born continuum solvent,^{34,35} however, the underlying conditional convergence of the Coulomb lattice summation was not addressed. Although fixed charge force fields may be designed for use with a spherical cutoff, this approach is based on the observation that the radial distribution function (RDF) of neat organic liquids asymptote to

unity at about one nanometer.³⁶ However, the RDFs for molecules within a periodic crystal do not decay to one but have periodic peaks at all lengths scales. Long-range correlations must be considered or inclusion of electrostatics can lead to systematic errors.^{14,22} Currently, refinement packages, such as CNS,^{10,37} Phenix,^{38,39} BUSTER,⁴⁰ and Refmac⁴¹ lack a rigorous Coulomb lattice summation method. Therefore, a key motivation for the current work was to create a state-of-the-art force field engine that can be incorporated into existing X-ray crystallography software and can handle data sets of any size—from small molecule crystals to ribosome crystals with millions of atoms.

We begin by describing the explicit incorporation of symmetry operators into PME electrostatics for the AMOEBA force field.⁸ Our algorithm consistently accommodates not only space groups but also replicates of the central unit cell and the combination of space group symmetry with replicates. Details of our parallelization scheme are presented for shared memory parallelization over CPU cores in combination with the option to offload the PME reciprocal space sum to a GPU coprocessor. Overall timings and the speed up relative to expansion to P1 for 30 crystals with a variety of space groups are discussed in order to demonstrate that for the first time PME electrostatics are affordable for X-ray refinement of all system sizes encountered in macromolecular crystallography. Finally, we compare rerefine-ment of 10 X-ray crystallography data sets with and without polarizable AMOEBA electrostatics.

II. PARTICLE MESH EWALD WITH SPACE GROUP SYMMETRY

A. Unit Cell, Space Group, and Asymmetric Unit Definitions. We define a lattice Λ in direct space by its basis vectors \mathbf{a} , \mathbf{b} , and \mathbf{c} that have Euclidean lengths a , b , and c , respectively. The Cartesian component of a vector will be denoted by a subscript, for example a_{α} , where $\alpha \in \{1,2,3\}$. The conjugate reciprocal lattice Λ^* is defined by basis vectors \mathbf{a}^* , \mathbf{b}^* , and \mathbf{c}^* that have Euclidean lengths a^* , b^* , and c^* , respectively. The unit cell U of the lattice Λ consists of all points \mathbf{r}_{frac} that have fractional coordinates with $0 \leq r_{\text{frac},\alpha} \leq 1$. Cartesian coordinates \mathbf{r} can be converted to fractional coordinates \mathbf{r}_{frac} via multiplication by a 3×3 fractionalization matrix $\mathbf{r}_{\text{frac}} = \mathbf{r}^t \mathbf{A}$, where the superscript t denotes the transpose of the Cartesian coordinate column vector into a row vector and the columns of \mathbf{A} are the reciprocal basis vectors. The inverse operation is given by $\mathbf{r} = \mathbf{r}_{\text{frac}}^t \mathbf{A}^{-1}$, where the rows of \mathbf{A}^{-1} are given by the direct basis vectors.

The space group of a crystal will be defined by its set of n_s symmetry operators. The i^{th} fractional symmetry operator ($\mathbf{R}_i, \mathbf{t}_i$) is composed of a 3×3 rotation matrix \mathbf{R}_i plus a translation vector \mathbf{t}_i . The analogous Cartesian symmetry operators will be denoted in *italics* as ($\mathbf{R}_i, \mathbf{t}_i$). We assume a chemical system that is modeled by a set of n_a unique atoms, which constitute the asymmetric unit. The position of each atom i is described by orthogonal coordinates \mathbf{r}_i and its permanent atomic charge, dipole, and quadrupole by $\{q_i, \mathbf{d}_i, \Phi_i\}$. The coordinates of any atom in the unit cell can then be generated by application of a symmetry operator to one of the atoms in the unique set. The atomic electrostatic moments also require application of the rotational part of the symmetry operator. To avoid unnecessary complexity and to keep the presentation as general as possible, discussion of the AMOEBA self-consistent field procedure that generates an induced dipole \mathbf{u}_i at each multipole site is restricted to the Supporting Information.

B. Replicates of the Unit Cell. Application of the minimum image convention to a unit cell whose smallest width is less than half the real space cutoff r_{cut} requires generation of atomic coordinates in replicates of the unit cell.⁴² The concept of symmetry operators can be generalized to a replicated super cell with $n_u = m_1 \times m_2 \times m_3$ copies of the unit cell arranged along scaled direct space basis vectors $\{\mathbf{a}_r = \mathbf{a}m_1, \mathbf{b}_r = \mathbf{b}m_2, \mathbf{c}_r = \mathbf{c}m_3\}$. The total number of symmetry operators for the replicated super cell is given by $n_r = n_s \times n_u$. The n_r fractional symmetry operators can be generated from the n_s fractional symmetry operators of the space group as

$$(\mathbf{R}_{ijkl}, \mathbf{t}_{ijkl}) = \left(\mathbf{R}_i, \left\{ \frac{t_{i,1} + j}{m_1}, \frac{t_{i,2} + k}{m_2}, \frac{t_{i,3} + l}{m_3} \right\} \right) \quad (2)$$

and the Cartesian symmetry operators generated by

$$(\mathbf{R}_{ijkl}, \mathbf{t}_{ijkl}) = (\mathbf{R}_i, \mathbf{t}_i + \mathbf{n}_{jkl}) \quad (3)$$

where $i = 1, \dots, n_s, j = 0, \dots, m_1 - 1, k = 0, \dots, m_2 - 1$ and $l = 0, \dots, m_3 - 1$. In both the fractional and Cartesian cases each replicates super cell rotation matrix is equal to a rotation matrix of the space group. The fractional translation vector of each symmetry operator is scaled down in proportion to the number of replicated unit cells in each dimension such that enumeration of the n_r unit cells over the indices j, k , and l fills the replicates super cell. Similarly, the lattice vector $\mathbf{n}_{jkl} = j\mathbf{a} + k\mathbf{b} + l\mathbf{c}$ is added to the original Cartesian translation vectors.

C. Electrostatics under Periodic Boundary Conditions. There are two distinct physical pictures associated with lattice summation that have subtle but important differences.¹⁴ The Ewald picture is based on an infinite lattice, which can be defined mathematically even though it is physically unrealizable. In this case the electrostatic potential obeys periodic boundary conditions, specifically $\Phi(\mathbf{r}) = \Phi(\mathbf{r} + n_1\mathbf{a} + n_2\mathbf{b} + n_3\mathbf{c})$ for any set of integers $\{n_1, n_2, n_3\}$. The second physical picture is based on embedding a finite spherical lattice of unit cells inside a continuum dielectric and then taking the limit as its radius is increased to infinity. In this case, the electrostatic potential does not obey periodic boundary conditions due to two additional fields. The first is proportional to the dipole moment of the unit cell Φ_{dipole} , and the second is due to the reaction field Φ_{RF} of the dielectric medium that is induced by the spherical lattice.^{43–46} If the dielectric of the surrounding medium is a vacuum, then there can be no reaction field, but the cell dipole field remains. On the other hand, if the dielectric of the medium is infinite, then the continuum reaction field cancels the dipole field. However, the physical picture of an embedded spherical lattice, even under so-called tinfoil boundary conditions with an infinite dielectric, is not equivalent to an infinite lattice and true periodic boundary conditions.¹²

We note that sampling from an embedded spherical lattice is conceptually problematic. Consider equivalent electrostatic charges separated by a lattice vector, for example, at locations \mathbf{r}_i and $\mathbf{r}_i + n_1\mathbf{a} + n_2\mathbf{b} + n_3\mathbf{c}$, that experience different dipole and/or reaction field forces. During a simulation only the coordinates of the central cell are explicitly propagated. In effect, the central unit cell and a unit cell on the edge of the embedded sphere are constrained to sample equivalent ensembles. Although both boundary conditions have limitations, in this work we focus on the Ewald infinite lattice picture and will not include further discussion of the embedded spherical lattice.

D. Asymmetric Unit Lattice Summation. To motivate the notation consider the electric potential at atom j located at \mathbf{r}_j due to a collection of n_c point charges around atom i located at \mathbf{r}_i , each with a magnitude and position denoted by c_k and \mathbf{r}_k :

$$V(\mathbf{r}_j) = \frac{1}{4\pi\epsilon_0} \sum_{k=1}^{n_c} \frac{c_k}{|\mathbf{r}_{ij} - \mathbf{r}_k|} \quad (4)$$

where $\mathbf{r}_{ij} = \mathbf{r}_j - \mathbf{r}_i$ and the Coulomb constant $1/4\pi\epsilon_0$ will be neglected for convenience throughout the rest of the article. The potential can be expanded in a Taylor series to give

$$V(\mathbf{r}_j) = \sum_{k=1}^{n_c} c_k \left(1 + r_{k,\alpha} \nabla_{i,\alpha} + \frac{1}{2} r_{k,\alpha} r_{k,\beta} \nabla_{i,\alpha} \nabla_{i,\beta} \right) \frac{1}{r_{ij}} \quad (5)$$

where $\nabla_{i,\alpha}$ is one component of the del operator acting at \mathbf{r}_i , $\alpha \in \{x, y, z\}$ and the Greek subscripts $\{\alpha, \beta, \gamma, \delta, \dots\}$ represent use of the Einstein summation convention for summing over tensor elements.⁴⁷ The monopole, dipole, and traceless quadrupole moments are defined as

$$\begin{aligned} q_i &= \sum_{k=1}^{n_c} c_k, \\ d_{i,\alpha} &= \sum_{k=1}^{n_c} c_k r_{k,\alpha}, \\ \Theta_{i,\alpha\beta} &= \sum_{k=1}^{n_c} c_k \left(\frac{3}{2} r_{k,\alpha} r_{k,\beta} - \frac{1}{2} r_k^2 \delta_{\alpha\beta} \right) \end{aligned} \quad (6)$$

where use of a traceless quadrupole is permitted since the potential satisfies the Laplace equation. Based on eq 6, we substitute the multipole moments back into the potential of eq 5 and define the multipolar operator L_i :

$$\begin{aligned} V(\mathbf{r}_j) &= L_i \left(\frac{1}{r_{ij}} \right), \\ L_i &= q_i + d_{i,\alpha} \nabla_{i,\alpha} + \frac{1}{3} \Theta_{i,\alpha\beta} \nabla_{i,\alpha} \nabla_{i,\beta} \end{aligned} \quad (7)$$

Similarly, we can define the potential at \mathbf{r}_i due to the multipole at \mathbf{r}_j using multipolar operator L_j :

$$\begin{aligned} V(\mathbf{r}_i) &= L_j \left(\frac{1}{r_{ij}} \right), \\ L_j &= q_j - d_{j,\alpha} \nabla_{i,\alpha} + \frac{1}{3} \Theta_{j,\alpha\beta} \nabla_{i,\alpha} \nabla_{i,\beta} \end{aligned} \quad (8)$$

where the sign difference between the multipolar operators is due to the relationship $\nabla_i = -\nabla_j$ for the function $|\mathbf{r}_i - \mathbf{r}_j|$. In the case of the AMOEBA force field, multipole coefficients are derived from electronic structure calculations on model chemical compounds using distributed multipole analysis (DMA).^{23,48,49}

The potential energy U of the n_a permanent multipoles that make up the asymmetric unit is given by the lattice summation:

$$U = \frac{1}{2} \frac{1}{n_s} \sum_n^* \sum_{s_i=1}^{n_s} \sum_{s_j=1}^{n_s} \sum_{i=1}^{n_a} \sum_{j=1}^{n_a} L_i(\mathbf{R}_{s_i}) L_j(\mathbf{R}_{s_j}) \frac{1}{|\mathbf{x}|} \quad (9)$$

where $\mathbf{x} = \mathbf{R}_{s_i} \mathbf{r}_i + \mathbf{t}_{s_i} - (\mathbf{R}_{s_j} \mathbf{r}_j + \mathbf{t}_{s_j}) + \mathbf{n}$, the outer sum is over all lattice vectors $\mathbf{n} = n_1\mathbf{a} + n_2\mathbf{b} + n_3\mathbf{c}$, the second and third sums are over the n_s symmetry operators of the space group that operate on sites i and j , respectively, and the inner sums are over the n_a multipole sites of the asymmetric unit. The asterisk denotes

skipping (or scaling) masked interaction pairs $(i, j) \in M$ in the list M and omission of self-interactions defined by $i = j$ for the central unit cell ($\mathbf{n} = 0$) and the identity symmetry operators ($s_i = s_j = 1$). A common example of masking is to omit the interaction between atoms that are covalently bonded. The multipolar operators L_i and L_j now include a Cartesian rotation matrix from a symmetry operator that rotates the multipole moments into the symmetry mate orientation:

$$L_i(\mathbf{R}) = q_i + (\mathbf{Rd}_i)_\alpha \nabla_{i,\alpha} + (\mathbf{R}\Theta_i \mathbf{R}^t)_{\alpha\beta} \nabla_{i,\alpha} \nabla_{i,\beta} \frac{1}{3} \quad (10)$$

and

$$L_j(\mathbf{R}) = q_j - (\mathbf{Rd}_j)_\alpha \nabla_{i,\alpha} + (\mathbf{R}\Theta_j \mathbf{R}^t)_{\alpha\beta} \nabla_{i,\alpha} \nabla_{i,\beta} \frac{1}{3} \quad (11)$$

Finally, division by two in eq 9 avoids double counting each interaction, and division by n_s converts from the unit cell energy to the asymmetric unit energy.

E. Asymmetric Unit Ewald Summation. Ewald summation¹² is based on multiplication of each term in eq 9 by a convergence function $\text{erfc}(\beta|\mathbf{x}|)$ and then by $1 - \text{erfc}(\beta|\mathbf{x}|) = \text{erf}(\beta|\mathbf{x}|)$ to give

$$\begin{aligned} U &= U_{\text{real}} + U_{\text{recip}} \\ &= \frac{1}{2} \frac{1}{n_s} \sum_n^* \sum_{s_i=1}^{n_s} \sum_{s_j=1}^{n_s} \sum_{i=1}^{n_a} \sum_{j=1}^{n_a} L_i(\mathbf{R}_{s_i}) L_j(\mathbf{R}_{s_j}) \frac{\text{erfc}(\beta|\mathbf{x}|)}{|\mathbf{x}|} \\ &\quad + \frac{1}{2} \frac{1}{n_s} \sum_n^* \sum_{s_i=1}^{n_s} \sum_{s_j=1}^{n_s} \sum_{i=1}^{n_a} \sum_{j=1}^{n_a} L_i(\mathbf{R}_{s_i}) L_j(\mathbf{R}_{s_j}) \frac{\text{erf}(\beta|\mathbf{x}|)}{|\mathbf{x}|} \end{aligned} \quad (12)$$

where β is the Ewald convergence parameter. The first summation U_{real} is rapidly decreasing and may be evaluated in real space by ignoring all terms outside of a cutoff radius r_c , which is typically chosen between 7 and 9 Å. An appropriate β can be determined by satisfying $\text{erfc}(\beta r_c)/r_c < \epsilon_{\text{real}}$ at the cutoff for a target error tolerance ϵ_{real} . The second term is smooth, periodic, and rapidly decreasing in reciprocal space if masked and if self-interactions are added back, which is discussed further below. The physical picture is that a 3D Gaussian charge density has been added and then subtracted at the location of each point charge (or appropriate gradients of the Gaussian density for dipole, quadrupole, or higher order moments). As the Ewald convergence parameter β is increased, for example, to satisfy the target error tolerance for a small real space cutoff, relatively higher frequencies must be included in the reciprocal space sum. In this manner β can be used to tune the relative rate of convergence of the two sums.

F. Real Space Summation. The real space sum in eq 12 can be simplified to

$$U_{\text{real}} = \frac{1}{2} \sum_{s_j=1}^{n_s} \sum_{i=1}^{n_a} \sum_{j=1}^{n_a} L_i(\mathbf{I}) L_j(\mathbf{R}_{s_j}) \frac{\text{erfc}(\beta|\mathbf{r}_i - (\mathbf{R}_{s_j} \mathbf{r}_j + \mathbf{t}_{s_j})|)}{|\mathbf{r}_i - (\mathbf{R}_{s_j} \mathbf{r}_j + \mathbf{t}_{s_j})|} \quad (13)$$

where the asterisk now indicates that $i = j$ interactions are neglected and masked interactions $(i, j) \in M$ are respected for the identity symmetry operator $s_j = 1$. A replicates super cell and n_r symmetry operators are required for a unit cell whose smallest width is less than half the real space cutoff.⁴² In this way all interactions within the real space cutoff are treated

consistently via application of the minimum image convention using the replicates super cell basis vectors $\{\mathbf{a}_r, \mathbf{b}_r, \mathbf{c}_r\}$. The sum over lattice vectors \mathbf{n} can be removed, since any lattice vector with length greater than zero produces interactions outside of the real space cutoff distance. When a replicates super cell is not required, then n_r is equal to n_s unit cell space group symmetry operators, and application of the minimum image convention is based on the unit cell basis vectors $\{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$. Since the real space energy for each copy of the asymmetric unit is equal, the sum over symmetry operators for multipole site i may be removed, and division by the number of space group operators is unnecessary.

We emphasize an important difference between eq 13 and the analogous eq (2.13) of Sagui et al. due to the inclusion of symmetry operators.¹⁶ The summations over i and j are reduced compared to a P1 unit cell from n_{p1} multipole sites to $n_a = n_{p1}/n_s$. If the asymmetric unit is large relative to the real space cutoff, then the reduction in terms relative to a calculation in P1 approaches a factor of n_s . When the real space work is numerically expensive relative to the reciprocal space work, as for multipolar force fields or electronic structure calculations, the speedup for the overall calculation approaches n_s .

Details for applying the multipolar operator in the case of the AMOEBA potential can be found in the Appendix of the work by Ren and Ponder and will not be repeated here.⁵ Briefly, the derivation of real space Ewald summation for multipoles presented by Smith⁵⁰ is modified by a Thole damping function⁵¹ at short range when computing polarization interactions for AMOEBA. However, we recommend consideration of the McMurchie–Davidson recursion,^{52,53} as presented by Sagui et al.¹⁶ when moments above quadrupole order are considered.

G. Reciprocal Space Summation. The reciprocal space summation requires adding and then subtracting the self-energy U_{self} and masked interaction energy U_{mask} that were excluded from eq 12 to give

$$U_{\text{recip}} = U_{\text{periodic}} - U_{\text{self}} - U_{\text{mask}} \quad (14)$$

The term U_{periodic} is smooth, periodic, and its Fourier transform is given by¹³

$$\hat{U}_{\text{periodic}} = \frac{1}{n_s} \frac{1}{2\pi V} \sum_{\mathbf{h} \neq 0} \frac{\exp(-\pi^2 s^2 / \beta^2)}{s^2} \hat{F}(\mathbf{h}) \hat{F}(-\mathbf{h}) \quad (15)$$

where $\mathbf{s} = \mathbf{h}^t \mathbf{A}^{-1}$ is the scattering vector, \mathbf{h} contains the Miller indices of a Bragg reflection, and $V = \mathbf{a} \cdot \mathbf{b} \times \mathbf{c}$ is the volume of the unit cell. The total multipolar structure factor,¹⁶ including a summation over unit cell symmetry operators, is given by

$$\hat{F}(\mathbf{h}) = \sum_{i=1}^{n_a} \sum_{s=1}^{n_s} \hat{L}_j(\mathbf{s}, \mathbf{R}_s) \exp[2\pi i \mathbf{h} \cdot (\mathbf{R}_s \mathbf{A}^t \mathbf{r}_j + \mathbf{t}_s)] \quad (16)$$

where the Fourier transform of the multipolar operator L_j is given by

$$\hat{L}_j(\mathbf{s}, \mathbf{R}) = q_j + 2\pi i (\mathbf{Rd}_j^t)_\alpha s_\alpha - 4\pi^2 (\mathbf{R}\Theta_j \mathbf{R}^t)_{\alpha\beta} s_\alpha s_\beta \quad (17)$$

Alternatively, symmetry operators can be applied in reciprocal space⁵⁴ to the structure factor for the asymmetric unit:

$$\hat{F}^A(\mathbf{h}) = \sum_{j=1}^{n_a} \hat{L}_j(\mathbf{s}, \mathbf{I}) \exp[2\pi i \mathbf{h} \cdot (\mathbf{r}_j^t \mathbf{A})] \quad (18)$$

based on the expression:

$$\hat{F}(\mathbf{h}) = \sum_{s=1}^{n_s} \hat{F}^A(\mathbf{R}_s^t \mathbf{h}) \exp(2\pi i \mathbf{h} \cdot \mathbf{t}_s) \quad (19)$$

where the term $\exp(2\pi i \mathbf{h} \cdot \mathbf{t}_s)$ is due to the translational part of the symmetry operator.

H. Ewald Self- and Masked Interactions. The self-interaction terms can be determined by taking the limit of $f(r) = \text{erf}(\beta r)/r$ and its partial derivatives, as specified by the multipolar operators in eq 15 in the limit $r \rightarrow 0$ to give

$$\begin{aligned} \lim_{r \rightarrow 0} f(r) &= \frac{2\beta}{\sqrt{\pi}} \\ \lim_{r \rightarrow 0} [\nabla_\alpha (-\nabla_\alpha) f(r)] &= \frac{4\beta^3}{3\sqrt{\pi}} \\ \lim_{r \rightarrow 0} [\nabla_\alpha^2 \nabla_\beta^2 (1/3)^2 f(r)] &= \frac{8\beta^5}{45\sqrt{\pi}} \\ \lim_{r \rightarrow 0} [\nabla_\alpha^4 (1/3)^2 f(r)] &= \frac{24\beta^5}{45\sqrt{\pi}} \end{aligned} \quad (20)$$

Based on the results of eq 20 the total self-interaction energy U_{self} that must be removed from U_{recip} is given by

$$U_{\text{self}} = \frac{1}{2} \sum_{i=1}^{n_a} \frac{2\beta}{\sqrt{\pi}} d_i^2 + \frac{4\beta^3}{3\sqrt{\pi}} d_{i,\alpha}^2 + \frac{16\beta^5}{45\sqrt{\pi}} \Theta_{i,\alpha\beta}^2 \quad (21)$$

which is consistent with the result of Aguado et al.⁵⁵ The quadrupole self-interaction term is based on intermediate steps that depend on it being traceless and symmetric. Since they have not been presented previously, we provide these steps below. From eq 20 the self-interaction for an element of the quadrupole trace is due to the interaction with itself and the other two trace elements:

$$\begin{aligned} &\frac{\beta^5}{45\sqrt{\pi}} [24\Theta_{\alpha\alpha}^2 + 8\Theta_{\alpha\alpha}(\Theta_{\beta\beta} + \Theta_{\gamma\gamma})] \\ &= \frac{\beta^5}{45\sqrt{\pi}} [24\Theta_{\alpha\alpha}^2 + 8\Theta_{\alpha\alpha}(-\Theta_{\alpha\alpha})] = \frac{16\beta^5}{45\sqrt{\pi}} \Theta_{\alpha\alpha}^2 \end{aligned} \quad (22)$$

while the self-interaction for an off-diagonal element is due to the interaction with itself and the symmetric element:

$$\frac{8\beta^5}{45\sqrt{\pi}} \Theta_{\alpha\beta}(\Theta_{\alpha\beta} + \Theta_{\beta\alpha}) = \frac{16\beta^5}{45\sqrt{\pi}} \Theta_{\alpha\beta}^2 \quad (23)$$

Masked terms $(i,j) \in M$ were easily accounted for in the real space sum but included at full strength in the Fourier sum to enforce exact periodicity. The overcounting can be removed by subtracting the real space sum over masked interactions within the asymmetric unit given by

$$U_{\text{mask}} = \frac{1}{2} \sum_{i,j \in M} L_i(\mathbf{I}) L_j(\mathbf{I}) \frac{\text{erf}(\beta |\mathbf{r}_i - \mathbf{r}_j|)}{|\mathbf{r}_i - \mathbf{r}_j|} \quad (24)$$

I. PME Reciprocal Space Summation. Instead of direct summation of structure factors, they can be computed via B-spline interpolation onto a discrete grid followed by 3D FFT. This is analogous to the method used to compute

crystallographic structure factors, with the notable differences that point multipoles are interpolated at grid points using B-splines that have finite support, whereas Gaussian form factors are explicitly evaluated at grid points and have infinite support necessitating truncation outside of a cutoff.⁵⁶ Smooth PME interpolates multipoles to a finite set of nearby grid points using cardinal B-splines $\theta_p(u)$ of order p as described originally by Essmann et al. for fixed charge models¹³ and later extended to higher order moments.^{15,16} The first order cardinal B-spline $\theta_1(u)$ is defined as the characteristic function of $[0,1]$ and higher orders recursively as the convolution product:

$$\theta_k = \theta_{k-1} * \theta_1 \quad (25)$$

The support of $\theta_k(u)$ is compact and given by $[0,k]$. The error of the PME approximation can be systematically reduced via higher order B-splines in tandem with finer grids.

The complex exponential in eq 16 may be expanded to

$$\exp(2\pi i \mathbf{h} \cdot \mathbf{u}_i) = \exp(2\pi i h_{u_i,1}) \exp(2\pi i k_{u_i,2}) \exp(2\pi i l_{u_i,3}) \quad (26)$$

where \mathbf{u}_i are the fractional coordinates of site i after application of the symmetry operator s_i as given by $\mathbf{u}_i = \mathbf{R}_s^t \mathbf{A}^t \mathbf{r}_i + \mathbf{t}_s$. The Euler exponential spline s_b is then used to interpolate each complex exponential¹³ of eq 26 as

$$\begin{aligned} \exp(2\pi i h_\alpha u_{i,\alpha}) &\approx s_b(h_\alpha, u_{i,\alpha}) \\ &= b_\alpha \left(\frac{h_\alpha}{N_\alpha} \right) \sum_{k=-\infty}^{\infty} \theta_p(N_\alpha u_{i,\alpha} - k) \\ &\quad \cdot \exp\left(2\pi i \frac{h_\alpha}{N_\alpha} k\right) \end{aligned} \quad (27)$$

for grid dimension N_α and coefficients $b_\alpha(h_\alpha/N_\alpha)$ given by

$$b_\alpha \left(\frac{h_\alpha}{N_\alpha} \right) = \frac{\exp[2\pi i(p-1)h_\alpha/N_\alpha]}{\sum_{k=0}^{p-2} \theta_p(k+1) \cdot \exp(2\pi i k h_\alpha/N_\alpha)} \quad (28)$$

The fractional grid array that includes the contributions of all multipoles is given by

$$Q(\mathbf{k}) = \sum_{s_i=1}^{n_s} \sum_{i=1}^{n_a} \sum_n \hat{L}_i(\mathbf{R}_{s_i}) \begin{bmatrix} \theta_p(u_{i,1}N_1 - k_1 - n_1N_1) \\ \times \theta_p(u_{i,2}N_2 - k_2 - n_2N_2) \\ \times \theta_p(u_{i,3}N_3 - k_3 - n_3N_3) \end{bmatrix} \quad (29)$$

where $\mathbf{k} = \{k_1, k_2, k_3\}$ is a point of the $\mathbf{N} = \{N_1, N_2, N_3\}$ sized 3D grid and $\mathbf{n} = \{n_1, n_2, n_3\}$ indicates a sum over all integer triples. The inner sum is actually finite due to local support of each B-spline. The discrete Fourier transform of eq 29 is given by

$$\hat{Q}(\mathbf{h}) = \sum_{k_1=0}^{N_1-1} \sum_{k_2=0}^{N_2-1} \sum_{k_3=0}^{N_3-1} Q(\mathbf{k}) \exp[2\pi i \mathbf{h} \cdot (k_1/N_1, k_2/N_2, k_3/N_3)] \quad (30)$$

and the analytic structure factor of eq 16 can be approximated as

$$\hat{F}(\mathbf{h}) = \begin{cases} B(\mathbf{h}) \hat{Q}(\mathbf{h}) & -\frac{1}{2} \leq \left\{ \frac{h}{N_1}, \frac{k}{N_2}, \frac{l}{N_3} \right\} < \frac{1}{2} \\ 0 & \text{otherwise} \end{cases} \quad (31)$$

where

$$B(\mathbf{h}) = b_1 \left(\frac{h}{N_1} \right) \cdot b_2 \left(\frac{k}{N_2} \right) \cdot b_3 \left(\frac{l}{N_3} \right) \quad (32)$$

The periodic portion of the reciprocal energy is then computed as before using eq 15. In our implementation the principle quantity of interest is the reciprocal electrostatic potential, and its gradients at each multipole site within the asymmetric unit, given by¹⁶

$$\begin{aligned} & \varphi_{\text{rec}}(\mathbf{r}_i) \\ &= \frac{1}{\pi V} \sum_{\mathbf{h} \neq 0} \frac{\exp(-\pi^2 \mathbf{h}^2 / \beta^2)}{\mathbf{h}^2} s_b(-h, u_{i,1}) s_b(-k, u_{i,2}) s_b(-l, u_{i,3}) \\ & \cdot \hat{F}(\mathbf{h}) = \sum_{\mathbf{n}} \theta_p(N_1 u_{i,1} - n_1) \theta_p(N_2 u_{i,2} - n_2) \theta_p(N_3 u_{i,3} - n_3) \\ & \cdot (G^*Q)(\mathbf{n}) \end{aligned} \quad (33)$$

where the second equality follows from Parseval's identity with Q given by eq 29 and G defined as the inverse discrete Fourier transform of a generalized influence function:

$$\begin{aligned} & \hat{G}(\mathbf{h}) \\ &= \begin{cases} \frac{1}{\pi V} |B(\mathbf{h})|^2 \frac{\exp(-\pi^2 s^2 / \beta^2)}{s^2} & -\frac{1}{2} \leq \left\{ \frac{h}{N_1}, \frac{k}{N_2}, \frac{l}{N_3} \right\} < \frac{1}{2}, \mathbf{s} \neq 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned} \quad (34)$$

The convolution of the pair potential G and multipole array Q gives the potential on the grid at \mathbf{n} , which is evaluated at a finite number of nonzero grid points in real space due to the finite support of the B-splines. The potential only needs to be evaluated for atoms within the asymmetric unit, which speeds up this part of the calculation by a factor of the number of space group symmetry operators. Gradients of the potential are found by taking gradients of θ_p , as described in earlier work.¹⁶

In reciprocal space, the convolution $C(\mathbf{n}) = (G^*Q)(\mathbf{n})$ becomes a simple multiplication for each structure factor:

$$\hat{C}(\mathbf{h}) = \begin{cases} \hat{G}(\mathbf{h}) \hat{Q}(\mathbf{h}) & -\frac{1}{2} \leq \left\{ \frac{h}{N_1}, \frac{k}{N_2}, \frac{l}{N_3} \right\} < \frac{1}{2} \\ 0 & \text{otherwise} \end{cases} \quad (35)$$

Performing the inverse discrete 3D FFT on \hat{C} generates the desired convolution product

$$\begin{aligned} C(\mathbf{n}) &= \frac{1}{N_1 N_2 N_3} \sum_{k_1=0}^{N_1-1} \sum_{k_2=0}^{N_2-1} \sum_{k_3=0}^{N_3-1} \hat{C}(\mathbf{k}) \\ & \exp[-2\pi i \mathbf{n} \cdot (k_1/N_1, k_2/N_2, k_3/N_3)] \end{aligned} \quad (36)$$

For optimal computational performance, it is advantageous to view the reciprocal space portion of the calculation in terms of a *single overall convolution operation*, rather than three serial steps as described above:

- (1) 3D FFT given in eq 30.
- (2) Reciprocal space multiplication given in eq 35.
- (3) 3D inverse FFT in eq 36. This idea will be emphasized in the following section on our parallel implementation.

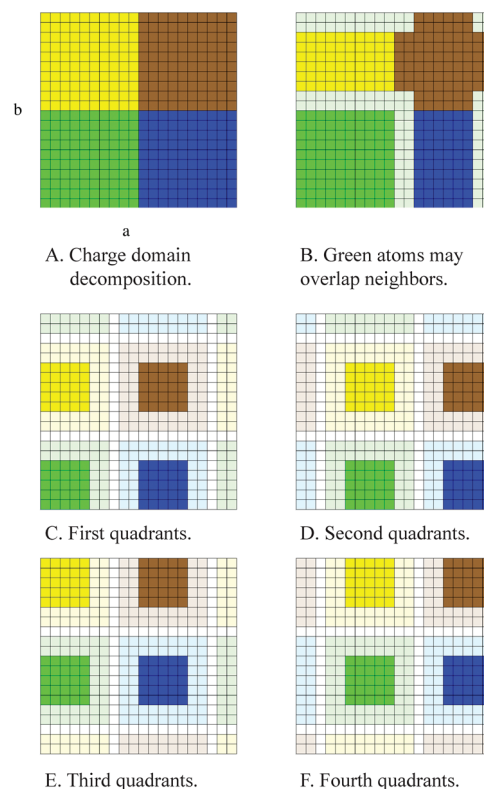


Figure 1. Panel A shows one face of a domain decomposition for a 20×20 grid of source density in fractional coordinates. All atoms within a color-coded domain are assigned to the same compute core. Panel B highlights in light green the grid region that is not assigned to the green core but receives source density from atoms in the green region (the effect of PBC is included). This is based on quintic b-Splines whose support region is a $5 \times 5 \times 5$ grid, although the method is easily adjusted to other support requirements. The yellow, brown, and blue cores must not attempt to modify the source density values at light-green grid points, while the green core is modifying them. Panels C–F demonstrate further subdivision of each CPU region into quadrants. In each of four synchronized steps, atoms within the active green, yellow, brown, and blue quadrant have their source density spread to the grid. Grid regions colored light green, light yellow, light brown, and light blue represent the maximum extent of source density spreading by their dark central quadrant. The white outline separating the maximum extent of support indicates that no two cores will require the same grid point during a step of the procedure. Slow atomic operations in software and hardware specific APIs are replaced by a few high-level thread synchronizations.

III. PARALLEL IMPLEMENTATION

We now focus on the shared memory parallelization of the reciprocal space portion of PME as implemented in FFX. This portion of the calculation is the limiting factor for force field energy evaluations for large biomolecular crystals due to $N \cdot \log(N)$ scaling of the FFT. The real space portion of our algorithm has also been parallelized, however, our view is that the combination of N -body summations with symmetry operators merits a separate, self-contained treatment. The reason is most zonal schemes or spatial decompositions assume nearly uniform particle density over the unit cell, which is not the case after removing redundant copies of the asymmetric unit.⁵⁷

First we discuss a general domain decomposition scheme for spreading source density onto the 3D FFT grid, as described by eq 29. Then we present two parallelization strategies for the

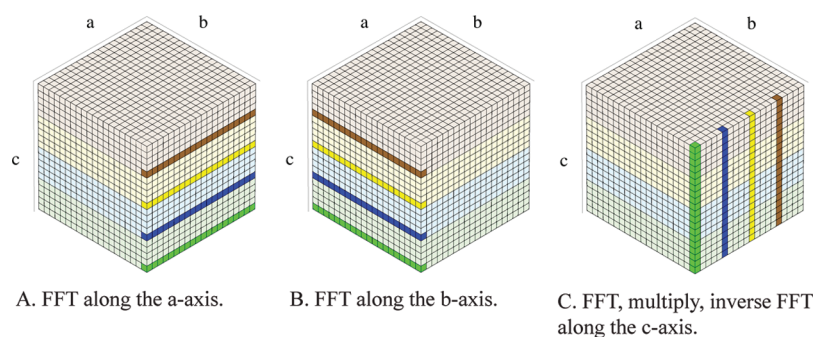


Figure 2. Panels A–C depict partitioning of the $20 \times 20 \times 20$ source grid into ab -planes that are distributed among four cores. Panels A and B represent 1D FFTs along the a -axis and b -axis, respectively. For these operations the required data is stored in memory assigned to the core doing the transform. Panel C shows the final 1D FFT that requires grid values that are distributed in memory across all cores. Before a thread moves on to the next row, the reciprocal space multiplications and inverse 1D FFT are performed to minimize cache misses or cache interference. This optimization is not possible for algorithms based on completing the 3D FFT for the entire grid before doing the reciprocal space multiplication. The final a -axis and b -axis inverse 1D FFTs are not shown.

Table 1. Shown Are Timings for Our 3D FFT Routines Based on Either Real or Complex Input Data for Transforms Sizes of 64^3 , 128^3 , and 256^3 ^a

method	64^3				128^3				256^3			
	1	8	X	CUDA	1	8	X	CUDA	1	8	X	CUDA
R	0.034	0.009	3.7		0.238	0.056	4.2		2.291	0.321	7.1	
R Conv.	0.028	0.006	4.4		0.228	0.040	5.6		2.140	0.248	8.6	
C	0.061	0.014	4.3	0.003	0.545	0.084	6.4	0.017	7.715	1.004	7.7	0.224
C Conv.	0.055	0.010	5.5		0.492	0.067	7.4		6.886	0.648	10.6	

^aThe real (R) and complex (C) timings are for the sum of 3D FFT and inverse 3D FFT called sequentially. For the real convolution (R Conv.) and complex convolution (C Conv.), the timing includes the 3D FFT, reciprocal space multiplication, and inverse 3D FFT treated a *single operation*. For each combination, both one and eight cores were tested, and the speed-up is shown in the column labeled X. The timings for CUDA are *italicized* because they were done in single precision, while all others calculations are double precision. All timings are in seconds. The CUDA library does not include *single operation* convolutions (R Conv. and C Conv.) and our work did not require implementation of the real CUDA sequential approach (R) so these table entries are blank. Speedups greater than 8 result from cache effects.

reciprocal space convolution. The key to the first algorithm is viewing the convolution as a single operation and has been parallelized for a shared memory JRE. The second algorithm is currently based on the nVidia CUDA language and its 3D FFT library, which necessitates viewing the convolution as a sequential series of three operations, as described above.

A. Spreading Source Density onto the 3D Grid. A key issue in parallelization of charge density spreading is that two threads of execution cannot be allowed to modify the value of any grid point concurrently. For example, consider the simplest possible spatial decomposition, namely two domains of equal size separated by two parallel planes (one plane is a periodic boundary and the other is parallel to it). Atoms that are near a plane will spread source density into both subdomains. Now consider dividing both subdomains in half by two additional planes, parallel to the first two, to give four total subdomains of equal size. As long as each subdomain dimension is as large as the support of the atomic source density, for example, five grid points in each dimension for quintic b-Splines, then it is guaranteed that subdomains that do not share a plane do not interpolate multipoles into each other. In our trivial example, therefore, threads may operate simultaneously on regions one and three without requiring access to the same grid point. When the two threads of execution complete regions one and three, they synchronously continue to regions two and four. More generally, there may be $d_\alpha = N_\alpha/b_\alpha$ subdomains and $p_\alpha = d_\alpha/2$ subdomain pairs along the

Table 2. Presented Are the Timings and Speed-up for the Evaluation of the Acetamide Crystal Structure Using the AMOEBA Force Field and PME Electrostatics^a

simulation cell	acetamide molecules	time (sec)	speed-up
$2 \times 2 \times 2$ P1	144	0.584	1.0
P1	18	0.114	5.1
H3c	1	0.016	36.5

^aThe $2 \times 2 \times 2$ replicated unit cell avoids the need for an explicit replicates algorithm, since the super cell edges are greater than twice the real space cutoff. The combination of space group and replicates operators in FFX gives a speed-up for the acetamide crystal of more than $36\times$ relative to the $2 \times 2 \times 2$ replicated unit cell without parallelization.

α -axis $\alpha \in \{a,b,c\}$ with grid dimension N_α and support requirement b_α . In two dimensions, there may be at most $d_{\alpha,\beta} = (N_\alpha/b_\alpha)(N_\beta/b_\beta)$ subdomains and $q_{\alpha,\beta} = d_{\alpha,\beta}/4$ subdomain quartets requiring 3 synchronization steps to avoid sharing planes, as shown in Figure 1. Finally, in three dimensions, there may be at most $d_{a,b,c} = (N_a/b_a)(N_b/b_b)(N_c/b_c)$ subdomains and $o_{a,b,c} = d_{a,b,c}/8$ octets requiring 7 synchronization steps to avoid sharing planes. Note that each division above must be done separately to ensure an even result along each axis.

B. Reciprocal Space Convolution. Parallelization of the reciprocal space convolution is of critical importance to the

parallel scaling of PME electrostatics. We briefly discuss our CPU parallelization scheme and its relative merits and also refer to more comprehensive and focused presentations.⁵⁸ The 3D transform is decomposed into 1D transforms along each axis, as shown in Figure 2. First N_b times N_c transforms of length N_a are performed along the a -axis. Then, $N_a \times N_c$ transforms of length N_b are performed along the b -axis. Finally, $N_a \times N_b$ transforms of length N_c are performed along the c -axis. If the data is packed in a 1D array in memory, with dimension a varying most quickly, dimension b varying second most quickly, and dimension c varying most slowly, then the transforms along the c -axis require the most severely nonlocal memory accesses. With this in mind, ab -planes are divided equally among available CPUs, and the first two sets of 1D transforms are very memory efficient. Transforms along the c -axis are then divided equally among available CPUs. Before each transform a thread-local array of length N_c is packed contiguously with values otherwise separated by $N_a \times N_b$ complex values in memory. For PME, the point-wise reciprocal multiply of eq 35 should be performed and the inverse FFT along the c -axis done immediately. Finally, the local result is copied back into the global array, before the thread moves on to its next c -axis transform.

Optionally, the reciprocal space calculation can be accelerated using the CUDA API, which does not include a convolution operation. Instead, the forward 3D FFT, reciprocal multiplication, and inverse 3D FFT are done sequentially. The CPUs still perform the real space calculation while the PME grid is transferred to the GPU, the convolution is performed, and finally the result is transferred back to main memory. The transfer time becomes insignificant for large 3D grids so that our overall algorithm scales $N \cdot \log(N)$ on the GPU (Table 1). Since we are currently using single precision for optional GPU acceleration, a discussion of single vs double precision is given in Section D of the Supporting Information.

IV. APPLICATIONS

There are a significant variety of applications where explicit inclusion of symmetry operators within PME described in this work and implemented within FFX may play an essential role. At the small organic molecule end of the spectrum is ab initio crystal structure prediction and solubility estimation.^{59,60} For example, the pharmaceutical industry is especially interested in polymorph prediction, where each polymorph can have different physical properties and bioavailability. Consider the case of acetaminophen, which crystallizes in three polymorphs.⁶¹ At the other end of the spectrum is the refinement of large macromolecular structures at low resolution where de novo model building can be problematic without previously determined high-resolution substructures.^{62,63} This work represents a first step toward demonstrating that the prior chemical information contained within the AMOEBA force field can be used to improve macromolecular models from refinement with medium- to low-resolution data sets.

A. Replicates for Small Molecule Crystals. Acetamide is an important model compound when developing a biomolecular force field and forms a crystal at standard temperature and pressure. For these reasons it was chosen to demonstrate the application of our space group plus replicates PME implementation of the AMOEBA force field to small organic crystals (Table 2). For comparison, a recent application of the fixed charge CHARMM force field to predict the crystal structure of *N*-(2-dimethyl-4,5-dinitrophenyl)

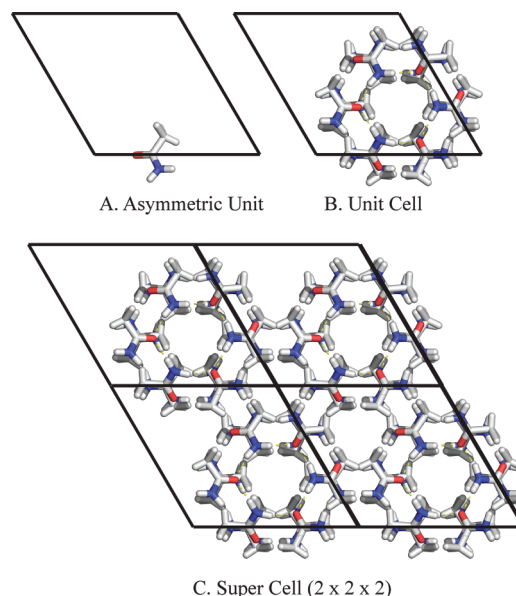


Figure 3. Shown in Panel A is the acetamide asymmetric unit (space group $H3c$). Panel B shows the unit cell after expansion to P1, which contains 18 molecules in identical chemical environments. Finally, Panel C shows a super cell ($2 \times 2 \times 2$) whose edge lengths are each at least twice the pair wise cutoff distance. Calculations of AMOEBA energies and gradients for the asymmetric unit are accelerated by a factor of more than $36\times$ relative to a code without replicates operators and $7\times$ relative to a code with replicates but without space group symmetry operators.

acetamide required expansion to P1 followed by the creation of at least two copies of the unit cell in each dimension.⁶⁴ Therefore, $2^3 \cdot N_{\text{symm}} - 1$ more copies of the asymmetric unit than necessary contribute to the computation of the asymmetric unit potential energy. In contrast, the algorithm presented here allows the calculation to be done on only the asymmetric unit. This is a useful step forward in terms of both force field quality (fixed charges replaced by polarizable atomic multipoles) and search efficiency. The small unit cell dimensions of the rhombohedral crystal (11.526, 11.526, 13.526, 90.0, 90.0, 120.0, space group $H3c$) are clearly not twice the distance of a converged real space cutoff for either PME or van der Waals energy. Since the unit cell contains 18 copies of acetamide, expansion to a $2 \times 2 \times 2$ P1 super cell contains 144 molecules, as shown in Figure 3.

B. Macromolecular X-ray Crystallography Refinement. Refinement of large macromolecular complexes, such as the ribosome, using a rigorous lattice summation method is a primary motivation for this work. For example, the importance of electrostatics to the mechanism of translation due to the interaction of divalent Mg^{2+} cations with the negatively charged phosphate backbone of rRNA, mRNA, and tRNA has been suggested by a recent 2.8 Å structure of *Thermus thermophilus*.⁶⁵ If the same structure could be solved to a higher resolution, perhaps below 2 Å, it is reasonable to expect further biological insights due to features that were unclear in the lower resolution electron density maps. As an example of such an improvement, consider the structure of the phenylalanine tRNA that was originally determined in the 1970s using ~ 3 Å resolution data sets⁶⁶ but was recently improved via a data set at 1.93 Å.⁶⁷ The new, higher resolution model exhibits six additional metal sites, an average change in torsional angles of $\sim 40^\circ$, alternate sugar puckers, and extensive differences in water structure.

Table 3. Shown Are Timings for the Evaluation of the AMOEBA Potential Energy for 30 Crystallography Data Sets^a

PDB	space	number of atoms	time (seconds)						speed-up
			AU	UC	AU				
					1 CPU	1 CPU	8 CPUs	8 CPUs + GPU	
ID	group	b_{symm}	AU	UC	1 CPU	1 CPU	8 CPUs	8 CPUs + GPU	speed-up
1AV1	P212121	4	13 224	52 896	53.2	28.5	3.6	2.0	27.1
1A7B	P212121	4	5928	23 712	13.2	5.4	1.0	0.9	14.3
1BL8	C2	4	5898	23 592	33.0	18.2	3.6	2.4	13.6
1DP0	P212121	4	76 805	307 220	235.2	95.9	20.9	14.7	16.0
1ISR	P3221	6	7067	42 402	23.6	11.6	1.8	1.1	20.6
1JL4	P4322	8	8749	69 992	36.1	15.2	3.1	2.0	17.8
1J5E	P41212	8	88 347	706 776	971.2	195.7	35.1	20.6	47.1
1PGF	I222	8	17 770	142 160	117.0	59.2	10.5	5.2	22.6
1RSU	I222	8	56 797	454 376	345.9	192.4	24.7	10.0	34.8
1XDV	P212121	4	25 155	100 620	68.0	29.5	6.8	5.4	12.5
1XXI	P212121	4	55 778	223 112	110.8	57.3	8.2	5.3	20.9
1X8W	P41212	8	31 220	249 760	324.0	63.2	9.1	5.0	64.5
1YE1	P21212	4	9366	37 464	16.9	6.6	1.2	0.9	18.0
1YI5	C2221	8	20 961	167 688	94.3	36.9	7.9	5.9	15.9
1Z9J	P4222	8	12 807	102 456	114.3	63.5	10.9	4.6	25.1
2A62	P4122	8	4911	39 288	26.6	15.8	3.0	1.5	18.3
2BF1	P43212	8	4853	38 824	28.2	12.8	2.8	1.7	16.7
2FNP	P21	2	4201	8402	5.5	3.4	0.7	0.6	9.4
2I36	P3112	6	15 266	91 596	61.8	31.8	5.0	3.1	19.6
2J00	P212121	4	487 164	1 948 656	3935.6	1513.7	200.6	85.1	46.2
2QAG	P4322	8	8947	71 576	76.5	62.2	8.3	3.8	20.0
2QUK	P622	12	6235	74 820	72.0	29.6	4.5	3.0	24.1
2R4R	C2	4	10 068	40 272	26.3	13.7	2.3	1.6	16.1
2VKZ	P43212	8	171,819	1,374,552	1036.8	398.8	47.6	20.3	51.2
3BBW	P61	6	8865	53 190	36.1	18.2	3.6	1.8	19.7
3CRW	P212121	4	8305	33 220	19.2	7.0	1.2	1.0	18.7
3DMK	I222	8	32 758	262 064	153.0	69.3	9.5	4.9	31.2
3DU7	P65	6	27 491	164 946	141.3	82.2	14.4	8.3	17.0
3FFN	P4212	8	22 645	181 160	114.7	44.8	7.5	4.8	23.8
3HN8	P41212	8	13 395	107 160	77.5	37.1	5.7	4.0	19.4
mean		6.3	42 093	239 798	278.9	107.3	15.5	7.7	24.1

^a The average number of atoms in the asymmetric unit (AU) is a 6.3 fold reduction compared to the average number of atoms in the unit cell (UC). The mean speed-up from space group symmetry, shared memory parallelization over 8 CPU cores, and a GPU coprocessor for the reciprocal space convolution is a factor of 24 \times .

We present timings for evaluation of the potential energy for 30 macromolecular crystals in Table 3. For example, the 3CRW asymmetric unit and unit cell are shown in Figure 4. The average number of atoms in the asymmetric unit (42 093) is already quite large relative to calculations that have been done with AMOEBA thus far.⁸ The average number of atoms in the unit cell after expansion to P1 symmetry (239 798) is 6.3 fold higher still. For these timings and the optimizations described below, the van der Waals cutoff was set to 9.0 Å. A polynomial switch was used to smoothly turn off the van der Waals potential energy over a window width of 0.9 Å (starting at 8.1 Å). For PME, the real space cutoff was set to 7.0 Å, the Ewald convergence parameter set to 0.545, the B-spline order to 5 and a reciprocal space grid density of 1.2 grid points per Å, which are the currently recommended values for use with AMOEBA.⁹ In some cases, the grid density was increased or decreased by no more than 10% to achieve power of 2 grid dimensions, which is currently

maximally efficient for the CUDA FFT library. In the future, we anticipate OpenCL FFT libraries that suffer less performance degradation for nonpower of two sizes. The AMOEBA self-consistent field (SCF) was converged to a tolerance of 0.01 RMS Debye. This is also known as the *mutual* polarization model. For high-temperature simulated annealing,⁶⁸ the accuracy of *mutual* polarization may be unnecessary, and a *direct* polarization approximation can be used instead. Under the *direct* polarization approximation, the total field of the permanent multipoles influences the polarizable sites but not the field of the induced dipoles themselves (for details see Section A of the Supporting Information). For this reason, the *direct* approximation is about an order of magnitude faster than the true AMOEBA potential that requires SCF convergence.

By using space group symmetry, shared memory parallelization, and a GPU coprocessor for the reciprocal space convolution, the average time for an energy evaluation of these large

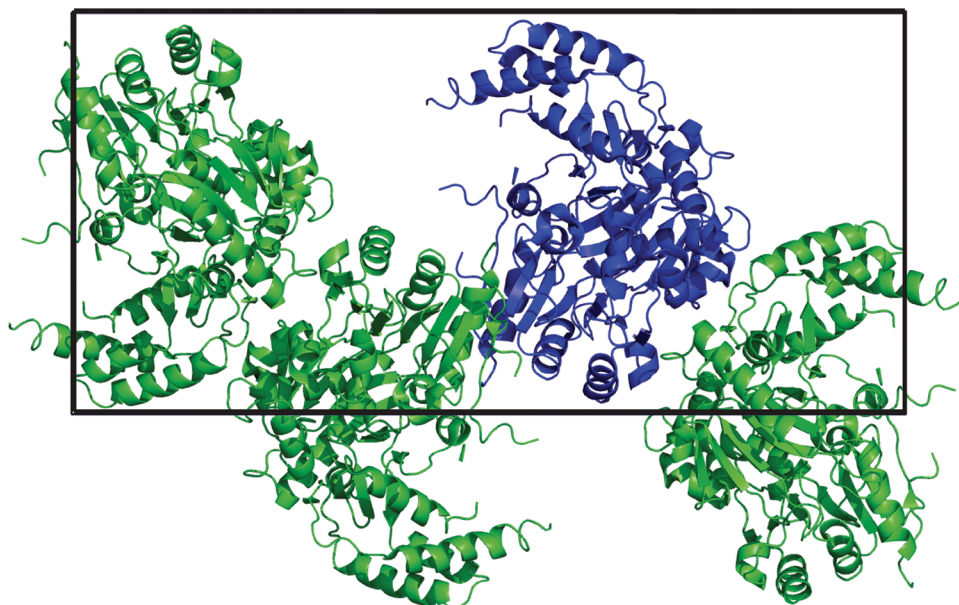


Figure 4. One of the smaller crystals (3CRW) is shown to give an impression of the number of real space interactions that are removed by considering only the asymmetric unit (blue) by eliminating redundant symmetry mates (green).

Table 4. Details for 10 Crystallography Data Used to Explore AMOEBA-Assisted Biomolecular X-ray Refinement Potential^a

model	res.	reported			FFX recalculated			geometry RMSD	
	(Å)	R	R _{free}	R _{free} - R	R	R _{free}	R _{free} - R	bonds (Å)	angles (°)
1A7B	3.1	23.9	30.6	6.7	25.9	31.1	5.2	0.005	1.27
1BL8	3.2	28.0	29.0	1.0	32.5	32.7	0.2	0.006	1.10
1DP0	1.7	15.7	21.1	5.4	16.1	20.3	4.2	0.018	2.92
1SFC	2.4	26.5	30.3	3.8	28.2	31.2	3.1	0.009	1.30
2FNP	2.6	28.5	30.8	2.3	27.2	29.0	1.8	0.010	1.90
2QUK	2.8	26.7	29.0	2.3	25.8	27.3	1.4	0.009	1.60
2R4R	3.4	21.7	27.0	5.3	23.4	28.0	4.6	0.007	1.40
3CRW	4.0	23.7	31.9	8.2	21.8	30.0	8.2	0.008	4.40
3FFN	3.0	22.2	27.1	4.9	21.6	26.6	5.0	0.012	1.40
3HN8	4.5	23.8	25.9	2.1	25.2	27.4	2.3	0.004	0.94
Mean	3.1	24.1	28.3	4.2	24.8	28.4	3.6	0.009	1.82

^a Including their resolution, R, R_{free}, R_{free} - R, and the RMS deviation (RMSD) of their bonds and angles from equilibrium values. Modest differences between the reported and re-calculated R, R_{free}, R_{free} - R are expected. Except for this table, relative differences in R, R_{free}, R_{free} - R due to the local optimization protocols are reported self-consistently using values calculated with FFX.

biomolecular crystals is reduced to 7.7 s on average for the AMOEBA potential using a desktop workstation, as shown in Table 3. For comparison, the same table is presented in Section B of the Supporting Information for the *direct* approximation to the AMOEBA potential, which reduces the average time to only 1.2 s. This illustrates that the most expensive part of the AMOEBA energy evaluation is the SCF rather than the permanent multipole electrostatics. For comparison, evaluation of the X-ray term of eq 1 for 1DP0 on 8 CPU cores costs a factor of 3 (~50 s) more than the force field term (15 s on 8 CPUs + GPU). Since the AMOEBA polarizable force field is generally less expensive or about equal to the X-ray term, the computational cost between the X-ray and force field terms is balanced.

A subset of 10 crystallography data sets, described in detail in Table 4, were selected from the 30 used for timings to compare the quality of biomolecular X-ray rerefinement with and without AMOEBA electrostatics. It is known that rerefinement from

deposited X-ray data with current methods including TLS treatment of B-factors in combination with a maximum likelihood X-ray target function improves most PDB entries.⁶⁹ We chose a broad resolution range, 1.7–4.5 Å with a mean of 3.1 Å, to promote more general conclusions. The averages of the originally reported R, R_{free}, and R_{free} - R are 24.1, 28.3, and 4.2%, respectively. The averages of R, R_{free}, and R_{free} - R recalculated using FFX gives 24.8, 28.4, and 3.6%, respectively, which provides a basis for self-consistent comparisons. Modest differences between the reported and recalculated R values are expected due to variations in the scattering engines of the various original refinement programs and FFX,²¹ such as their treatment of bulk solvent and crystal anisotropy. All R values in Table 5 were calculated in FFX.

Beginning with the ten PDB structures listed in Table 4, a stringent local optimization procedure was applied using the refinement target given by eq 1 where the X-ray refinement term

Table 5. Refinement Statistics and MolProbity Metrics for 10 Biomolecular Crystallography Data Sets^a

model	structure	R	R_{free}	$R_{\text{free}} - R$	clashscore	poor	Ramachandran (%)		MolProbity
						rot. (%)	outliers	avored	score
1A7B	PDB	25.9	31.1	5.2	24.9	8.3	0.8	95.1	2.93
3.1	vdW	20.5	31.6	11.1	1.4	14.1	0.3	92.6	2.20
	AMOEBA	20.6	30.9	10.3	4.4	12.2	0.0	97.0	1.86
1BL8	PDB	32.5	32.7	0.2	80.9	23.1	4.2	72.9	4.23
3.2	vdW	24.7	30.4	5.8	1.4	11.4	4.2	81.3	2.39
	AMOEBA	24.7	29.2	4.5	4.6	7.9	3.2	90.5	2.08
1DP0	PDB	16.1	20.3	4.2	10.6	3.6	0.2	96.5	2.20
1.7	vdW	14.7	18.9	4.2	3.7	2.1	0.1	97.3	1.53
	AMOEBA	14.9	19.0	4.1	3.8	1.9	0.1	97.3	1.51
1SFC	PDB	28.2	31.2	3.1	48.5	7.9	1.0	95.1	3.18
2.4	vdW	22.5	30.6	8.1	3.1	10.0	1.1	94.5	2.23
	AMOEBA	23.1	30.6	7.6	5.0	7.0	0.5	97.2	1.90
2FNP	PDB	27.2	29.0	1.8	75.0	9.8	5.8	82.5	3.80
2.6	vdW	21.0	30.4	9.4	2.4	15.4	2.5	87.1	2.54
	AMOEBA	21.1	28.3	7.2	5.5	13.3	2.1	92.5	2.34
2QUK	PDB	25.8	27.3	1.4	43.8	13.9	3.5	88.4	3.58
2.8	vdW	21.9	30.1	8.2	6.5	12.7	2.7	86.8	2.82
	AMOEBA	22.5	30.1	7.6	8.8	13.9	2.4	90.6	2.76
2R4R	PDB	23.4	28.0	4.6	80.3	11.3	4.4	79.1	3.92
3.4	vdW	20.5	27.1	6.6	4.3	13.6	4.6	81.3	2.79
	AMOEBA	20.9	26.7	5.8	7.1	12.5	2.4	87.3	2.66
3CRW	PDB	21.8	30.0	8.2	70.8	9.1	4.8	76.8	3.82
4.0	vdW	18.1	30.2	12.1	0.4	9.3	2.1	84.8	2.03
	AMOEBA	19.9	29.8	9.9	1.0	8.7	1.3	89.9	1.90
3FFN	PDB	21.6	26.6	5.0	13.1	10.9	1.6	94.2	2.81
3.0	vdW	16.9	25.2	8.3	1.8	11.2	1.4	92.9	2.19
	AMOEBA	17.1	24.5	7.4	3.4	10.1	1.5	95.6	2.01
3HN8	PDB	25.2	27.4	2.3	32.5	9.3	1.9	87.6	3.34
3.5	vdW	19.3	24.7	5.5	5.5	19.0	2.2	85.9	2.91
	AMOEBA	19.5	24.9	5.4	7.2	16.3	2.5	89.3	2.78
mean	PDB	24.8	28.4	3.6	48.0	10.7	2.8	86.8	3.38
	vdW	20.0	27.9	7.9	3.0	11.9	2.1	88.5	2.36
	AMOEBA	20.4	27.4	7.0	5.1	10.4	1.6	92.7	2.18

^a R and R_{free} for the starting models (PDB) were re-calculated using FFX to enable self-consistent comparisons. Local optimizations were performed, as described in the text, without electrostatics (vdW) and using the AMOEBA polarizable force field (AMOEBA) for the chemical term of eq 1. The mean improvements in R_{free} are 0.5 and 1.0% under vdW and AMOEBA protocols, respectively. AMOEBA shows the greatest reduction in poor side-chain rotamers and Ramachandran backbone outliers (outliers). In addition, AMOEBA achieves the greatest increase in favorable backbone (θ, ϕ) dihedral pairs (avored) and overall MolProbity score. Although the vdW protocol achieves a lower clashscore than AMOEBA, this is due to incorrect treatment of weak hydrogen bonds (C–H···O) by this heuristic.

was described by Fenn et al.²¹ and the chemistry term is a simplification of the AMOEBA potential chosen to mimic the REPEL force field used in CNS.¹⁰ Specifically, both electrostatics and torsional terms were turned off to give a nonbonded force that only included van der Waals interactions (this potential is referred to by the abbreviation vdW below). The X-ray weight (w_a) was set to 2.5 based on optimization of the mean R_{free} for the 10 data sets following 10 rounds of coordinate and B-factor optimization under the vdW potential (data not shown), although a weight in the range of 1.0–5.0 does not change our conclusions. The coordinate optimizations during each round were converged to a RMS gradient of 0.05 kcal/mol/Å, and the B-factor optimizations were converged to a RMS gradient of 0.005 (unitless). Beginning from the final vdW model, further

optimization was performed using eq 1 based on either the full AMOEBA force field (referred to as AMOEBA) or the *direct* polarization approximation to AMOEBA (results in Section C of the Supporting Information).

The final models from the vdW and AMOEBA local optimization protocols will first be compared based on R_{free} , $R_{\text{free}} - R$, and local structural metrics computed using MolProbity,^{70,71} as shown in Table 5. The AMOEBA optimization reduced R_{free} by an average of 0.5% relative to the starting models obtained via the vdW optimization procedure. In addition, inclusion of electrostatics in the AMOEBA optimizations reduced overfitting ($R_{\text{free}} - R$) by 0.9%. It is important to emphasize that although the R, R_{free} and $R_{\text{free}} - R$ reported in Table 5 are calculated self-consistently in FFX, comparisons between the deposited

Table 6. Geometric Statistics, Coordinate Superposition RMSDs and the Relative Energy per Residue for 10 Biomolecular Crystallography Data Sets^a

model		geometry RMSD		coord. RMSD (Å)		
res (Å)	potential	bond (Å)	angle (°)	C _α	heavy	rel. energy/residue (kcal/mol)
1A7B	vdW	0.014	2.64	0.39	0.67	0.0
3.1	AMOEBA	0.013	2.77	0.43	0.77	−20.0
1BL8	vdW	0.017	3.08	0.60	0.87	0.0
3.2	AMOEBA	0.016	3.27	0.74	0.98	−7.3
1DP0	vdW	0.020	2.94	0.11	0.24	0.0
1.7	AMOEBA	0.020	3.07	0.11	0.25	−16.4
1SFC	vdW	0.014	2.74	0.35	0.62	0.0
2.4	AMOEBA	0.014	3.07	0.42	0.75	−9.1
2FNP	vdW	0.016	2.76	0.44	0.82	0.0
2.6	AMOEBA	0.016	3.06	0.50	0.96	−9.9
2QUK	vdW	0.015	2.79	0.30	0.56	0.0
2.8	AMOEBA	0.015	3.04	0.30	0.59	−6.8
2R4R	vdW	0.015	2.65	0.60	0.90	0.0
3.4	AMOEBA	0.014	2.94	0.64	0.98	−6.7
3CRW	vdW	0.014	2.39	0.59	0.92	0.0
4.0	AMOEBA	0.014	2.83	0.92	1.24	−12.0
3FFN	vdW	0.015	2.68	0.31	0.48	0.0
3.0	AMOEBA	0.014	2.95	0.31	0.52	−7.0
3HN8	vdW	0.016	2.82	0.44	0.68	0.0
3.5	AMOEBA	0.016	3.12	0.45	0.73	−8.4
mean	vdW	0.016	2.75	0.41	0.67	0.0
	AMOEBA	0.015	3.01	0.48	0.78	−10.4

^a The structures used here correspond to those of Table . Compared to the vdW structures, inclusion of electrostatics slightly decreased the bond RMSD from equilibrium but increased the angle RMSD. The C_α coordinate RMSD, computed relative to the starting PDB structures, was 0.41 and 0.48 Å under the vdW and AMOEBA protocols, respectively. The mean heavy atom coordinate RMSD was 0.67 and 0.78 Å for vdW and AMOEBA, respectively. Finally, the AMOEBA potential energy per residue, relative to the vdW structure, was reduced by 10.4 kcal/mol by optimization with AMOEBA polarizable electrostatics.

structures and vdW minima are not significant in terms of drawing conclusions with respect to the merits of a potential energy function. What is significant, however, is the reduction in R_{free} and overfitting upon local optimization from the baseline vdW minima using the full AMOEBA model, with all other adjustable parameters fixed. We also note the increase in the average $R_{\text{free}} - R$ in going from the deposited PDB structures to the vdW minima. This is explained by the original structures being refined without hydrogens and/or not being optimized to a tight convergence criterion.

The vdW optimization drastically reduced the van der Waals clashscore from a mean of 48.0 to only 3.0, which is the number of van der Waals clashes per 1000 atoms. Similar improvements can be achieved via the all-hydrogen force field in the initial³⁷ and more recent¹⁰ versions of CNS. We note that formation of energetically favorable weak hydrogen bonds (i.e., C–H···O) that are driven by electrostatics are incorrectly counted as clashes by MolProbity (among other simplifications). This explains why the AMOEBA protocol causes a modest increase in clashscore relative to the vdW potential and points out a limitation of the generally useful MolProbity clashscore heuristic.

Although the percentage of poor side-chain rotamers was increased by the vdW optimization from a mean of 10.7 to 11.9%, the backbone Ramachandran statistics improved. Specifically, the percentage of outliers was reduced from 2.8 to 2.1% and the percentage of favorable (ϕ, φ) dihedral pairs increased from 86.8

to 88.5%. Unlike the vdW result, inclusion of electrostatics in the AMOEBA protocol slightly reduced the percentage of poor side-chain rotamers to 10.4%. Backbone Ramachandran outliers were further reduced from the vdW result to 1.6% under AMOEBA, and favorable (ϕ, φ) dihedral pairs were further improved from the vdW result by 4.2%.

The overall MolProbity score is a log-weighted combination of the clashscore, percentage of bad side-chain rotamers, and the unfavored Ramachandran percentage that indicates the crystallographic resolution at which those values would be expected.⁷¹ The mean value of the starting models is 3.38, which is slightly worse than the actual average crystallographic resolution of 3.1 Å. Under the vdW and AMOEBA protocols, the score was improved to 2.36 and 2.18, respectively. Therefore, MolProbity judges the quality of the AMOEBA structures to be consistent with a mean crystallographic resolution that is 0.92 Å better than the actual mean of the 10 data sets. Without going into details, the contribution from the clashscore was fixed to the vdW result when calculating MolProbity scores for AMOEBA (and direct) to ameliorate limitations of the clashscore heuristic for weak hydrogen bonds.

The RMS deviation of the bonds and angles from equilibrium values for the optimized structures referred to in Table 5 are listed in Table 6. Relative to the starting models, the bond RMSD increased from 0.009 to 0.015 Å and the angle RMSD from 1.822 to 2.749°. The increase of approximately 50% in both cases may

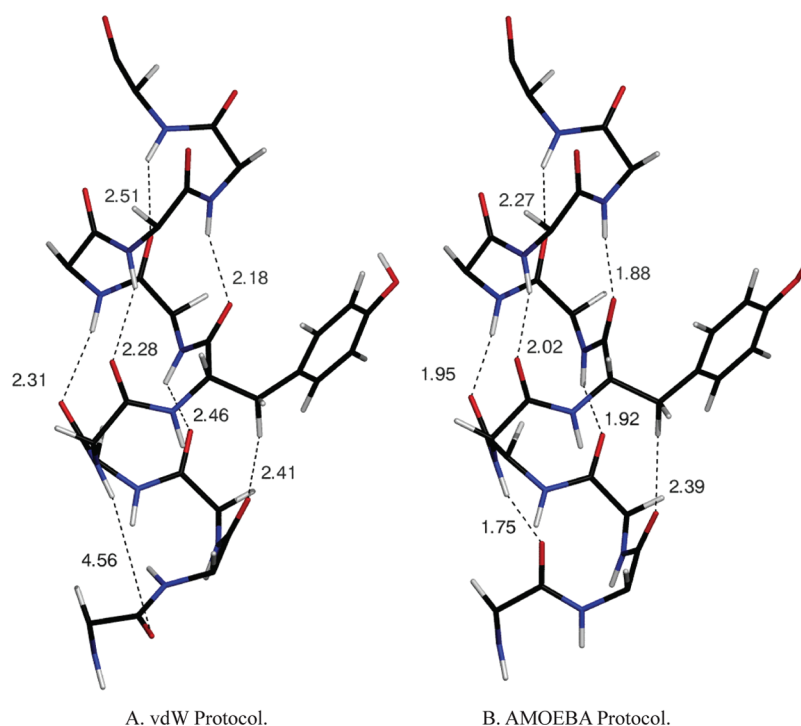


Figure 5. Panels A and B show residues A214–A224 of the human β_2 adrenergic G-protein-coupled receptor (2R4R) after optimization with the vdW and AMOEBa protocols, respectively. Canonical i to $i + 4$ α -helix hydrogen-bonding distances are explicitly drawn. The vdW protocol (Panel A) exhibits five potential canonical hydrogen bonds, but the mean N–H \cdots O distance (2.35 Å) is relatively large. The AMOEBa protocol (Panel B) lengthened the helix by one residue relative to the vdW protocol. These six α -helix hydrogen bonds show a mean N–H \cdots O distance (1.97 Å) in the optimal energetic range. We note that the second AMOEBa oxygen from the bottom of the image forms a weak hydrogen bond to the $i + 4$ C_β –H instead of to the $i + 4$ N–H, which could be improved by replacing the local optimization protocol used in this work with a global one.

be explained by the relatively stiff bond and angle force constants often used by crystallographic refinement software compared to AMOEBa force constants that are fit to vacuum vibrational frequencies measured experimentally or determined by electronic structure calculations. Alternatively, the increase could be due to differences in the weighting (w_a) between the X-ray and chemical terms in the target function. The mean energy stored in the bonds and angles of the final AMOEBa structures was only 0.08 and 0.14 kcal/mol, respectively, which is easily less than kT at the range of temperatures crystallographic data is collected (100–300 K).

The RMS deviation in the atomic coordinates for C_α atoms and for all heavy atoms from the PDB starting structures is also shown in Table 6. The vdW and AMOEBa protocols moved C_α atoms by an average of 0.41 and 0.49 Å, respectively. Larger RMS coordinate deviations of 0.67 and 0.78 Å were observed for all heavy atoms after the vdW and AMOEBa protocols, respectively. A relative potential energy per residue was calculated by subtracting the asymmetric unit potential energy of the AMOEBa model from that of the vdW model using the full AMOEBa potential energy function (AMOEBa is a better estimate of the true potential energy than the vdW potential) followed by division by the number of residues to achieve an estimate independent of protein size. The relatively small local perturbation of the coordinates due to optimization with electrostatics nonetheless resulted in a lowering of the relative potential energy per residue by more than 10 kcal/mol, which is comparable to a protein/drug binding free energy (for example, benzamidine binding to trypsin is favorable by 6.3–7.3 kcal/mol).⁷ The dramatic energetic improvement is consistent with electrostatic

stabilization from the formation of hydrogen bonds, as shown in Figure 5 for an α -helix of the human β_2 adrenergic G-protein-coupled receptor (2R4R, 3.4 Å). The AMOEBa protocol lengthened the α -helix by one residue relative to the vdW protocol, which is consistent with the 1.0 Å higher resolution 2RH1 structure (2RH1, 2.4 Å, Figure S-1 in the Supporting Information).

V. CONCLUSIONS

From this work we conclude the AMOEBa polarizable force field evaluated with PME electrostatics is capable of improving macromolecular X-ray crystallography refinement starting from structures obtained by optimization with only van der Waals nonbonded forces. The improvements lower R_{free} by 0.5%, reduce overfitting by 0.9% and increase the number of residues with favorable backbone conformations by 4.2%. This is consistent with electrostatics driving local conformational shifts toward favorable hydrogen-bonding networks, especially for repetitive secondary structure, as shown in Figure 5. The mean MolProbity score for our final models suggests geometric quality consistent with a mean crystallographic resolution of 2.18 Å, which is 0.92 Å better than the true mean of 3.1 Å.

We have shown that PME electrostatics benefits from the explicit incorporation of space group symmetry to reduce both memory and CPU demands. Further acceleration was achieved using shared memory parallelization and a GPU coprocessor for the $N \cdot \log(N)$ reciprocal space convolution of the PME algorithm. The X-ray crystallography refinements were carried out in FFX, which currently depends on v. 1.6 of the JRE and v. 3.1 of

the CUDA API for additional acceleration using a GPU coprocessor. Relative to a single CPU core after expansion to P1, the combination of space group symmetry, shared memory parallelization over 8 Intel Xeon E5530 CPU cores at 2.4 GHz, and a Tesla M1060 GPU coprocessor at 1.30 GHz showed an average speed-up of more than 24× for large macromolecular crystals that average 240 000 atoms in the unit cell.

One limitation of our results is the lack of a physical treatment of bulk solvent, such as Poisson–Boltzmann⁷² (PB) or generalized Kirkwood⁷³ (GK) continuum electrostatics. Currently, the PB and GK models for AMOEBA do not include explicit support for symmetry operators or periodic boundary conditions, but it should be possible to extend them in this respect. Although the cost of numerical solutions to the PB equation for AMOEBA is prohibitive for macromolecular X-ray refinement, it may be possible to combine the analytic GK approximation with PME. It has been noted previously that global optimization via simulated annealing may lead to unreasonable side-chain conformations without continuum solvent, especially for charged residues at the surface of a macromolecule that incorrectly experience a vacuum environment.³² Although the addition of a continuum solvent³³ followed by global optimization via simulated annealing⁶⁸ has been suggested, the fundamental problem of how to combine analytic continuum electrostatics with a rigorous lattice summation method remains an open question.

Although it is beyond the scope of this work, it is of interest to compare the improvements in model quality from AMOEBA electrostatics evaluated with PME to refinement using fixed charge electrostatics evaluated with spherical cutoffs as in CNS¹⁰ or to electronic structure methods.^{74,75} For example, the general features of the hydrogen-bonding network in Figure 5 might be reproduced by refinement with a fixed charge force field. However, the advantages of the polarizable AMOEBA model over fixed charge potentials have already been studied in detail for the structural properties of water,^{5,6} ion solvation thermodynamics,⁷⁶ protein–ligand binding affinities,^{7,77} and small molecule structural and thermodynamic observables.⁸ Similar advantages have also been observed for the CHARMM polarizable force field based on the classical drude oscillator.^{78–81}

■ ASSOCIATED CONTENT

S Supporting Information. The definition of the AMOEBA *direct* polarization approximation and self-consistent field procedure. Timings and refinements for the test systems using direct polarization are then presented. Analysis of single precision and double precision force accuracy in the context of macromolecular X-ray refinement. One additional figure is presented that demonstrates AMOEBA-assisted refinement of 2R4R (3.4 Å). This information is available free of charge via the Internet at <http://pubs.acs.org/>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: michael.schnieders@gmail.com. Telephone: (650) 995-3526.

■ ACKNOWLEDGMENT

The authors acknowledge Pengyu Ren, Thomas A. Darden, Alan M. Grossfield, and Jay W. Ponder for the PME code in TINKER this work began from and for helpful discussions. The

authors also wish to thank Axel T. Brunger for suggestions with regards to formulating self-consistent tests of refinement target functions. This work has been supported by an award from the NSF to Vijay S. Pande, Jay W. Ponder, Teresa Head-Gordon, and Martin Head-Gordon for “Collaborative Research: Cyberinfrastructure for Next Generation Biomolecular Modeling” (award number CHE-0535675) and by the Howard Hughes Medical Institute. For computer resources we acknowledge NSF award CNS-0619926 that supports the Bio-X2 cluster.

■ REFERENCES

- (1) Schnieders, M. J.; Fenn, T. D.; Pande, V. S.; Brunger, A. T. Polarizable atomic multipole X-ray refinement: application to peptide crystals. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2009**, *65* (9), 952–965.
- (2) Fenn, T. D.; Schnieders, M. J.; Brunger, A. T.; Pande, V. S. Polarizable atomic multipole X-ray refinement: hydration geometry and application to macromolecules. *Biophys. J.* **2010**, *98* (12), 2984–2992.
- (3) Fenn, T. D.; Schnieders, M. J.; Mustyakimov, M.; Wu, C.; Langan, P.; Pande, V. S.; Brunger, A. T. Reintroducing electrostatics into macromolecular crystallographic refinement: application to neutron crystallography and DNA hydration. *Structure* **2011**, *19*.
- (4) Ren, P.; Ponder, J. W. Consistent treatment of inter- and intramolecular polarization in molecular mechanics calculations. *J. Comput. Chem.* **2002**, *23* (16), 1497–1506.
- (5) Ren, P.; Ponder, J. W. Polarizable atomic multipole water model for molecular mechanics simulation. *J. Phys. Chem. B* **2003**, *107* (24), 5933–5947.
- (6) Ren, P.; Ponder, J. W. Temperature and pressure dependence of the AMOEBA water model. *J. Phys. Chem. B* **2004**, *108* (35), 13427–13437.
- (7) Jiao, D.; Golubkov, P. A.; Darden, T. A.; Ren, P. Calculation of protein–ligand binding free energy by using a polarizable potential. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105* (17), 6290–6295.
- (8) 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.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T. Current status of the AMOEBA polarizable force field. *J. Phys. Chem. B* **2010**, *114*, 2549–2564.
- (9) Ponder, J. W. *TINKER: Software Tools for Molecular Design*, 5.0; Jay W. Ponder: Saint Louis, MO, 2009.
- (10) Brunger, A. T., Version 1.2 of the Crystallography and NMR system. *Nature Protocols* **2007**, *2*, (11), 2728–2733.
- (11) 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.
- (12) Ewald, P. P. Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Annalen der Physik* **1921**, *369* (3), 253–287.
- (13) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle-mesh Ewald method. *J. Chem. Phys.* **1995**, *103* (19), 8577–8593.
- (14) Sagui, C.; Darden, T. A. Molecular dynamics simulations of biomolecules: long-range electrostatic effects. *Annu. Rev. Biophys. Biomol. Struct.* **1999**, *28*, 155–179.
- (15) Toukmaji, A.; Sagui, C.; Board, J.; Darden, T. Efficient particle-mesh Ewald based approach to fixed and induced dipolar interactions. *J. Chem. Phys.* **2000**, *113* (24), 10913–10927.
- (16) Sagui, C.; Pedersen, L. G.; Darden, T. A. Towards an accurate representation of electrostatics in classical force fields: Efficient implementation of multipolar interactions in biomolecular simulations. *J. Chem. Phys.* **2004**, *120* (1), 73–87.
- (17) Shan, Y. B.; Klepeis, J. L.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. Gaussian split Ewald: A fast Ewald mesh method for molecular simulation. *J. Chem. Phys.* **2005**, *122* (5), 13.
- (18) Cerutti, D. S.; Case, D. A. Multi-level Ewald: A hybrid multi-grid/fast Fourier transform approach to the electrostatic particle-mesh problem. *J. Chem. Theory Comput.* **2010**, *6* (2), 443–458.

- (19) Cerutti, D. S.; Duke, R. E.; Darden, T. A.; Lybrand, T. P. Staggered mesh ewald: an extension of the smooth particle-mesh Ewald method adding great versatility. *J. Chem. Theory Comput.* **2009**, *5* (9), 2322–2338.
- (20) Neelov, A.; Holm, C. Interlaced P3M algorithm with analytical and ik-differentiation. *J. Chem. Phys.* **2010**, *132* (23), 15.
- (21) Fenn, T. D.; Schnieders, M. J.; Brunger, A. T. A smooth and differentiable bulk-solvent model for macromolecular diffraction. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2010**, *66* (9), 1024–1031.
- (22) Karttunen, M.; Rottler, J.; Vattulainen, I.; Sagui, C. Electrostatics in biomolecular simulations: Where are we now and where are we heading? In *Computational Modeling of Membrane Bilayers*; Elsevier Academic Press Inc.: San Diego, CA, 2008; Vol. 60, pp 49–89.
- (23) Stone, A. J.; Alderton, M. Distributed multipole analysis-methods and applications. *Mol. Phys.* **1985**, *56* (5), 1047–1064.
- (24) Stone, A. J. Intermolecular potentials. *Science* **2008**, *321* (5890), 787–789.
- (25) Ponder, J. W.; Case, D. A. Force fields for protein simulations. In *Advances in Protein Chemistry*; Academic Press: San Diego, CA, 2003; Vol. 66, pp 27–85.
- (26) Harder, E.; Anisimov, V. M.; Vorobyov, I. V.; Lopes, P. E. M.; Noskov, S. Y.; MacKerell, A. D.; Roux, B. Atomic level anisotropy in the electrostatic modeling of lone pairs for a polarizable force field based on the classical Drude oscillator. *J. Chem. Theory Comput.* **2006**, *2* (6), 1587–1597.
- (27) Rafat, M.; Popelier, P. L. A. A convergent multipole expansion for 1,3 and 1,4 Coulomb interactions. *J. Chem. Phys.* **2006**, *124* (14), 7.
- (28) Rafat, M.; Shaik, M.; Popelier, P. L. A. Transferability of quantum topological atoms in terms of electrostatic interaction energy. *J. Phys. Chem. A* **2006**, *110* (50), 13578–13583.
- (29) Solano, C. J. F.; Pendas, A. M.; Francisco, E.; Blanco, M. A.; Popelier, P. L. A. Convergence of the multipole expansion for 1,2 Coulomb interactions: The modified multipole shifting algorithm. *J. Chem. Phys.* **2010**, *132* (19), 10.
- (30) Afonine, P. V.; Grosse-Kunstleve, R. W.; Adams, P. D.; Lunin, V. Y.; Urzhumtsev, A. On macromolecular refinement at subatomic resolution with interatomic scatterers. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2007**, *63*, 1194–1197.
- (31) Dawson, B., The covalent bond in diamond. *Proc. R. Soc. London, Ser. A* **1967**, *298*, (1454), 264–288.
- (32) Weis, W. I.; Brunger, A. T.; Skehel, J. J.; Wiley, D. C. Refinement of the influenza-virus hemagglutinin by simulated annealing. *J. Mol. Biol.* **1990**, *212* (4), 737–761.
- (33) Moulinier, L.; Case, D. A.; Simonson, T. Reintroducing electrostatics into protein X-ray structure refinement: bulk solvent treated as a dielectric continuum. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2003**, *59*, 2094–2103.
- (34) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. *J. Phys. Chem. A* **1997**, *101* (16), 3005–3014.
- (35) Bashford, D.; Case, D. A. Generalized Born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129–152.
- (36) 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* (2), 926–935.
- (37) Brunger, A. T.; Adams, P. D.; Clore, G. M.; DeLano, W. L.; Gros, P.; Grosse-Kunstleve, R. W.; Jiang, J.-S.; Kuszewski, J.; Nilges, M.; Pannu, N. S.; Read, R. J.; Rice, L. M.; Simonson, T.; Warren, G. L. Crystallography & NMR System: A new software suite for macromolecular structure determination. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1998**, *54*, 905–921.
- (38) Adams, P. D.; Grosse-Kunstleve, R. W.; Hung, L. W.; Ioerger, T. R.; McCoy, A. J.; Moriarty, N. W.; Read, R. J.; Sacchettini, J. C.; Sauter, N. K.; Terwilliger, T. C. PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2002**, *58*, 1948–1954.
- (39) Adams, P. D.; Afonine, P. V.; Bunkoczi, G.; Chen, V. B.; Davis, I. W.; Echols, N.; Headd, J. J.; Hung, L.-W.; Kapral, G. J.; Grosse-Kunstleve, R. W.; McCoy, A. J.; Moriarty, N. W.; Oeffner, R.; Read, R. J.; Richardson, D. C.; Richardson, J. S.; Terwilliger, T. C.; Zwart, P. H. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2010**, *66* (2), 213–221.
- (40) Bricogne, G.; Blanc, E.; Brandl, M.; Flensburg, C.; Keller, P.; Paciorek, P.; Roversi, P.; Sharff, A.; Smart, O.; Vonrhein, C.; Womack, T. BUSTER, 2.9; Global Phasing Ltd.: Cambridge, U.K., 2010.
- (41) Bailey, S., The CCP4 suite - programs for protein crystallography. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1994**, *50*, 760–763.
- (42) Smith, W., The minimum image convention in non-cubic MD cells. *CCPS Information Quarterly* **1989**, *30*, (35).
- (43) Smith, E. R. Electrostatic energy in ionic crystals. *Proc. R. Soc. London, Ser. A* **1981**, *375* (1763), 475–505.
- (44) Deleew, S. W.; Perram, J. W.; Smith, E. R. Simulation of electrostatic systems in periodic boundary conditions. 1. Lattice sums and dielectric constants. *Proc. R. Soc. London, Ser. A* **1980**, *373* (1752), 27–56.
- (45) Deleew, S. W.; Perram, J. W.; Smith, E. R. Simulation of electrostatic systems in periodic boundary conditions. 2. Equivalence of boundary conditions. *Proc. R. Soc. London, Ser. A* **1980**, *373* (1752), 57–66.
- (46) Deleew, S. W.; Perram, J. W.; Smith, E. R. Simulation of electrostatic systems in periodic boundary conditions. 3. Further theory and applications. *Proc. R. Soc. London, Ser. A* **1983**, *388* (1794), 177–193.
- (47) Stone, A. J. *The Theory of Intermolecular Forces*: Clarendon Press: Oxford, 1996; Vol. 32, p 264.
- (48) Stone, A. J. Distributed multipole analysis: Stability for large basis sets. *J. Chem. Theory Comput.* **2005**, *1* (6), 1128–1132.
- (49) Shi, Y.; Wu, C.; Ponder, J. W.; Ren, P. Multipole Electrostatics in Hydration Free Energy Calculations. *J. Chem. Comput.* **2010**, *32*.
- (50) Smith, W. Point multipoles in the Ewald summation (revisited). *CCPS Information Quarterly* **1982**, *4*, 13.
- (51) Thole, B. T. Molecular polarizabilities calculated with a modified dipole interaction. *Chem. Phys.* **1981**, *59* (3), 341–350.
- (52) McMurchie, L. E.; Davidson, E. R. One- and two-electron integrals over Cartesian Gaussian functions. *J. Comput. Phys.* **1978**, *26* (2), 218–231.
- (53) Challacombe, M.; Schwegler, E.; Almlof, J. Recurrence relations for calculation of the Cartesian multipole tensor. *Chem. Phys. Lett.* **1995**, *241* (1–2), 67–72.
- (54) Brunger, A. A memory-efficient fast Fourier transformation algorithm for crystallographic refinement on supercomputers. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1989**, *45*, 42–50.
- (55) Aguado, A.; Madden, P. A. Ewald summation of electrostatic multipole interactions up to the quadrupolar level. *J. Chem. Phys.* **2003**, *119* (14), 7471–7483.
- (56) Agarwal, R. C. New least-squares refinement technique based on fast Fourier-transform algorithm. *Acta Crystallogr., Sect. A: Cryst. Phys., Diff., Theor. Gen. Crystallogr.* **1978**, *34*, 791–809.
- (57) Bowers, K. J.; Dror, R. O.; Shaw, D. E. Zonal methods for the parallel execution of range-limited N-body simulations. *J. Comput. Phys.* **2007**, *221* (1), 303–329.
- (58) Frigo, M.; Johnson, S. G. The design and implementation of FFTW3. *Proc. IEEE* **2005**, *93* (2), 216–231.
- (59) Neumann, M. A.; Leusen, F. J. J.; Kendrick, J. A major advance in crystal structure prediction. *Angew. Chem., Int. Ed.* **2008**, *47* (13), 2427–2430.
- (60) Day, G. M.; Cooper, T. G.; Cruz-Cabeza, A. J.; Hejczyk, K. E.; Ammon, H. L.; Boerrigter, S. X. M.; Tan, J. S.; Della Valle, R. G.; Venuti, E.; Jose, J.; Gadre, S. R.; Desiraju, G. R.; Thakur, T. S.; van Eijck, B. P.; Facelli, J. C.; Bazterra, V. E.; Ferraro, M. B.; Hofmann, D. W. M.; Neumann, M. A.; Leusen, F. J. J.; Kendrick, J.; Price, S. L.; Misquitta, A. J.; Karamertzanis, P. G.; Welch, G. W. A.; Scheraga, H. A.; Arnautova,

Y. A.; Schmidt, M. U.; van de Streek, J.; Wolf, A. K.; Schweizer, B. Significant progress in predicting the crystal structures of small organic molecules - a report on the fourth blind test. *Acta Crystallogr., Sect. B: Struct. Crystallogr. Cryst. Chem.* **2009**, *65*, 107–125.

(61) Perrin, M. A.; Neumann, M. A.; Elmaleh, H.; Zaske, L. Crystal structure determination of the elusive paracetamol Form III. *Chem. Commun.* **2009**, *22*, 3181–3183.

(62) Davies, J. M.; Brunger, A. T.; Weis, W. I. Improved structures of full-length p97, an AAA ATPase: Implications for mechanisms of nucleotide-dependent conformational change. *Structure* **2008**, *16* (5), 715–726.

(63) Brunger, A. T.; DeLaBarre, B.; Davies, J. M.; Weis, W. I. X-ray structure determination at low resolution. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2009**, *65*, 128–133.

(64) Bazterra, V. E.; Thorley, M.; Ferraro, M. B.; Facelli, J. C. A distributed computing method for crystal structure prediction of flexible molecules: An application to N-(2-dimethyl-4,5-dinitrophenyl) acetamide. *J. Chem. Theory Comput.* **2007**, *3* (1), 201–209.

(65) Selmer, M.; Dunham, C. M.; Murphy, F. V.; Weixlbaumer, A.; Petry, S.; Kelley, A. C.; Weir, J. R.; Ramakrishnan, V. Structure of the 70S ribosome complexed with mRNA and tRNA. *Science* **2006**, *313* (5795), 1935–1942.

(66) Robertus, J. D.; Ladner, J. E.; Finch, J. T.; Rhodes, D.; Brown, R. S.; Clark, B. F. C.; Klug, A. Structure of yeast phenylalanine tRNA at 3 Å resolution. *Nature* **1974**, *250* (5467), 546–551.

(67) Shi, H. J.; Moore, P. B. The crystal structure of yeast phenylalanine tRNA at 1.93 Å resolution: A classic structure revisited. *RNA* **2000**, *6* (8), 1091–1105.

(68) Brunger, A. T. Simulated annealing in crystallography. *Annu. Rev. Phys. Chem.* **1991**, *42*, 197–223.

(69) Joosten, R. P.; Womack, T.; Vriend, G.; Bricogne, G. Refinement from deposited X-ray data can deliver improved models for most PDB entries. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2009**, *65*, 176–185.

(70) Davis, I. W.; Leaver-Fay, A.; Chen, V. B.; Block, J. N.; Kapral, G. J.; Wang, X.; Murray, L. W.; Arendall, W. B.; Snoeyink, J.; Richardson, J. S.; Richardson, D. C. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* **2007**, *35*, W375–W383.

(71) Chen, V. B.; Arendall, W. B.; Headd, J. J.; Keedy, D. A.; Immormino, R. M.; Kapral, G. J.; Murray, L. W.; Richardson, J. S.; Richardson, D. C. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2009**, *66*, 12–21.

(72) Schnieders, M. J.; Baker, N. A.; Ren, P. Y.; Ponder, J. W. Polarizable atomic multipole solutes in a Poisson-Boltzmann continuum. *J. Chem. Phys.* **2007**, *126*, 12.

(73) Schnieders, M. J.; Ponder, J. W. Polarizable atomic multipole solutes in a generalized Kirkwood continuum. *J. Chem. Theory Comput.* **2007**, *3* (6), 2083–2097.

(74) Yu, N.; Li, X.; Cui, G. L.; Hayik, S. A.; Merz, K. M. Critical assessment of quantum mechanics based energy restraints in protein crystal structure refinement. *Protein Sci.* **2006**, *15* (12), 2773–2784.

(75) Li, X.; Hayik, S. A.; Merz, K. M. QM/MM X-ray refinement of zinc metalloenzymes. *J. Inorg. Biochem.* **2010**, *104*, 512–522.

(76) Grossfield, A.; Ren, P. Y.; Ponder, J. W. Ion solvation thermodynamics from simulation with a polarizable force field. *J. Am. Chem. Soc.* **2003**, *125* (50), 15671–15682.

(77) Jiao, D.; Zhang, J. J.; Duke, R. E.; Li, G. H.; Schnieders, M. J.; Ren, P. Y. Trypsin-ligand binding free energies from explicit and implicit solvent simulations with polarizable potential. *J. Comput. Chem.* **2009**, *30* (11), 1701–1711.

(78) Lopes, P. E. M.; Roux, B.; MacKerell, A. D. Molecular modeling and dynamics studies with explicit inclusion of electronic polarizability: theory and applications. *Theor. Chem. Acc.* **2009**, *124* (1–2), 11–28.

(79) Lopes, P. E. M.; Lamoureux, G.; Roux, B.; MacKerell, A. D. Polarizable empirical force field for aromatic compounds based on

the classical drude oscillator. *J. Phys. Chem. B* **2007**, *111* (11), 2873–2885.


(80) Lamoureux, G.; Roux, B. Absolute hydration free energy scale for alkali and halide ions established from simulations with a polarizable force field. *J. Phys. Chem. B* **2006**, *110* (7), 3308–3322.

(81) Lamoureux, G.; Harder, E.; Vorobyov, I. V.; Roux, B.; MacKerell, A. D. A polarizable model of water for molecular dynamics simulations of biomolecules. *Chem. Phys. Lett.* **2006**, *418* (1–3), 245–249.

From Coarse Grained to Atomistic: A Serial Multiscale Approach to Membrane Protein Simulations

Phillip J. Stansfeld and Mark S.P. Sansom*

Department of Biochemistry, University of Oxford, South Parks Road, Oxford, OX1 3QU, United Kingdom

 Supporting Information

ABSTRACT: Coarse-grained molecular dynamics provides a means for simulating the assembly and the interactions of membrane protein/lipid complexes at a reduced level of representation, allowing longer and larger simulations. We describe a fragment-based protocol for converting membrane simulation systems, comprising a membrane protein embedded in a phospholipid bilayer, from coarse-grained to atomistic resolution, for further refinement and analysis via atomistic simulations. Overall, this provides a method for generating an accurate and well equilibrated membrane protein/lipid complex. We exemplify the protocol using the acid-sensing/amiloride-sensitive ion channel protein (ASIC) channel protein, a trimeric integral membrane protein. The method is further evaluated using a test set of 10 different membrane proteins of differing size and complexity. Simulations are assessed in terms of protein conformational drift, lipid/protein interactions, and lipid dynamics.

INTRODUCTION

Membrane proteins play key roles in cell biology, e.g., in transport and in signaling. As a consequence, membrane proteins account for ~25% of genes¹ and ~50% of the potential drug targets.² There is ongoing progress in the determination of membrane protein structures by X-ray diffraction and other methods,³ which has resulted in ~250 unique structures (see http://blanco.biomol.uci.edu/membrane_proteins_xtal.html for a summary). However, such structures only occasionally (e.g., for Aqp0)^{4,5} reveal full details of protein/lipid interactions. At the same time spectroscopic^{6,7} and functional⁸ studies indicate the importance of characterizing the nature of the interactions of membrane proteins with their lipid bilayer environment.

Molecular dynamics (MD) simulations and related methods have an important role in helping us to fully understand the structural dynamics of membrane proteins.^{9,10} However, prior to commencing these simulations the lipid/protein system should first be near optimally configured. The standard computational method for incorporating a protein into a lipid bilayer is to position the protein within the preformed membrane, delete the overlapping lipids, and then equilibrate the resulting complex.¹¹ However, this method requires prior knowledge of the trans-membrane region of the protein, with regions of high hydrophobicity, flanked by tyrosine, tryptophan, and basic residues used as indicators.¹² A number of online tools such as the Orientations of Proteins in Membranes (OPM) database¹³ (<http://opm.phar.umich.edu/>) can be used to guide this process, but they simplify the representation of the lipid bilayer to a hydrophobic slab.

A number of studies have shown that coarse-grained molecular dynamics (CG-MD) simulations¹⁴ may be used to characterize the interactions of membrane proteins with bilayer lipids^{15,16} (<http://sbcb.bioch.ox.ac.uk/cgdb>). However, this method simplifies the representation of both protein and lipid and so inevitably, e.g., the energetics of protein/lipid interactions are approximated.^{17,18} Other approaches include use of a mixed

CG-AT system (e.g., refs 19–22). It is therefore desirable to be able to adopt a serial multiscale approach,²³ whereby CG-MD simulations may be used to efficiently explore membrane protein/lipid interactions, yielding system configuration which may then be converted to atomistic resolution and further refined and characterized in detail by atomistic MD (AT-MD) simulations. One challenge in undertaking such an approach is to develop robust and efficient procedures for conversion of complex protein/lipid bilayer systems from CG to AT representations. This can be achieved in a number of ways.^{24,25} Here we describe a fragment-based approach, which we evaluate via application to a test set of 10 different membrane proteins and use as the basis of a comparison of lipid/protein interactions as predicted by CG and AT-MD simulations.

METHODS

CG to Atomistic Conversion. The overall problem is to generate an atomistic protein/lipid bilayer system which retains all of the key interactions of the corresponding coarse-grained model. Coarse-grained to atomistic (CG2AT) conversion is an intrinsically under determined problem, and so additional stereochemical information must be added, directly or indirectly. In our approach this additional information is provided by fragment-based libraries. For each protein a CG protein/lipid bilayer system complex, generated by a self-assembly CG-MD simulation,^{16,26} was used as the starting point for the conversion process. Thus, each system had previously been subjected to 500 ns of CG-MD to allow the self-assembly and equilibration of a lipid bilayer around the protein. The conversion starts by renaming all CG lipid and protein particles to their atomistic counterpart. The protocol (written in perl) uses a number of tools from Gromacs v4.5.3,²⁷ with both the standard Martini

Received: October 4, 2010

Published: March 16, 2011

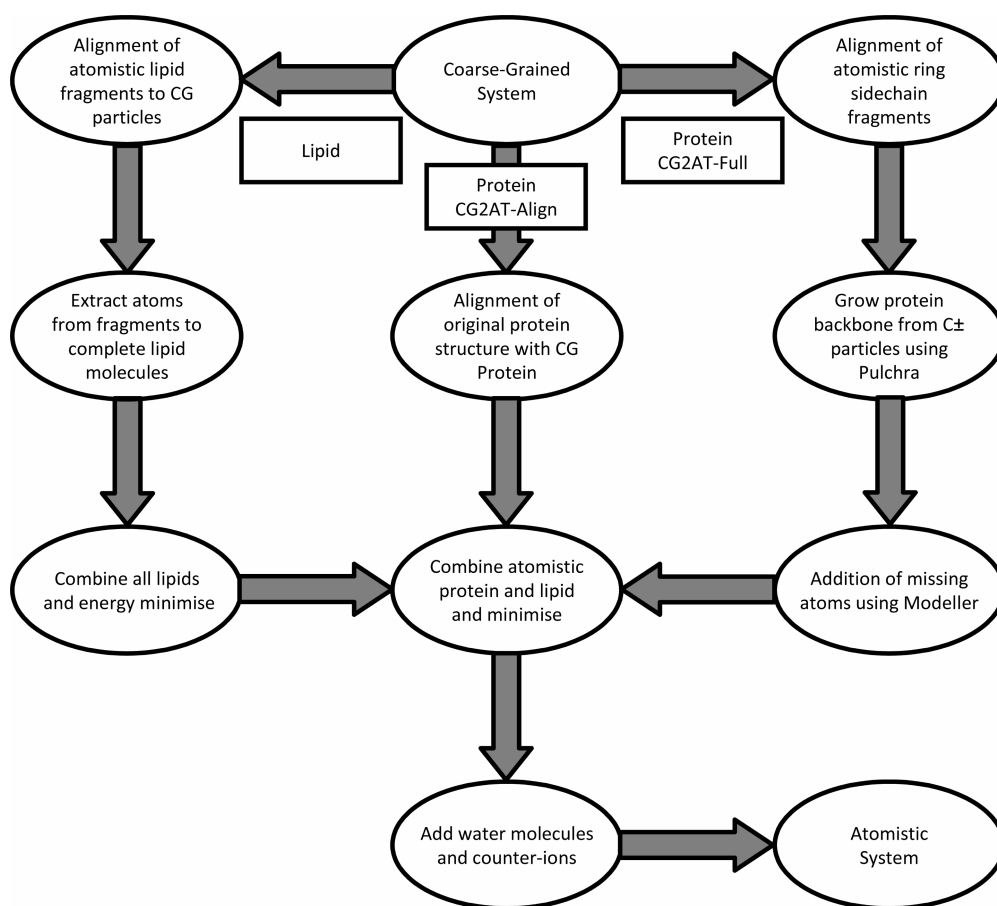


Figure 1. Flowchart describing the CG to AT conversion. Description of the CG2AT conversion methodology. Lipids and proteins are converted independently. The protein may be converted by either realignment of the original atomistic PDB structure (CG2AT-align) or reconstruction of the atomistic coordinates from the CG model using Pulchra and Modeller (CG2AT-full).

v2.1^{28,29} CG model and our local modification of Martini¹⁸ being supported. The protocol currently supports GROMOS, OPLS, and CHARMM36 force fields for the atomistic simulations. A flowchart detailing the conversion process is shown in Figure 1. A typical conversion takes 15 min on a standard Linux workstation for a 25 000 CG-particle system, with relatively small timing differences depending on the options used.

Protein Conversion. Two alternative approaches were used for CG2AT conversion of the protein. The first (CG2AT-full) uses Modeller and Pulchra to construct an atomistic protein structure from the CG particles.^{30,31} The second (CG2AT-align) structurally aligns the original protein PDB structure on the CG model of the protein. The decision as to which to use depends on whether or not one wishes to include any protein conformational changes that may have occurred during the CG-MD simulation. Thus, CG2AT-full would be used if one wished to carry over a (limited) protein conformational change occurring in the protein during the CG-MD step, e.g., due to the interaction of a nonmembrane domain with lipid headgroups. In contrast, one might better use CG2AT-align if CG-MD simulations were being used for a relatively rigid membrane protein simply to establish an optimal lipid bilayer environment as a starting point for extensive AT-MD simulations. Ultimately, which approach to use depends on the nature of the specific questions being addressed in a given simulation study.

In CG2AT-full the backbone of the protein is grown from the CG C α particles using the Pulchra algorithm.³¹ To guide the reconstruction of the aromatic side chains, fragments of ring structures are first aligned to the CG particles. The missing side chain atoms are then added using the complete_pdb function in Modeller;³⁰ this method preserves the original coordinates from the CG particles (Supporting Information, Figure S1). Subsequent conjugate gradients energy minimization is then applied using Modeller to reduce the internal steric clashes of the model. The protein is then energy minimized further through 500 steps of steepest descents using Gromacs. The stereochemical quality of the generated models was evaluated using Procheck.³²

In CG2AT-align the original atomistic structure used as the starting point for the CG simulations is structurally aligned with the protein CG particles and then energy minimized using 500 steps of steepest descents. There is also an option to align based only on the transmembrane region of the protein (selected using a consensus of sequence-based TM helix predictions).³³ This latter option allows for an improved alignment of the transmembrane region of the protein, especially for proteins with large, mobile nonmembrane domains.

Lipid Conversion. Lipids were converted by alignment of atomistic lipid fragments (Figure 2) to the CG particles of each lipid molecule. This is repeated for all lipids in the system, which is then subjected to 5000 steps of steepest descents energy minimization. For PG and PS containing lipids, a further limited

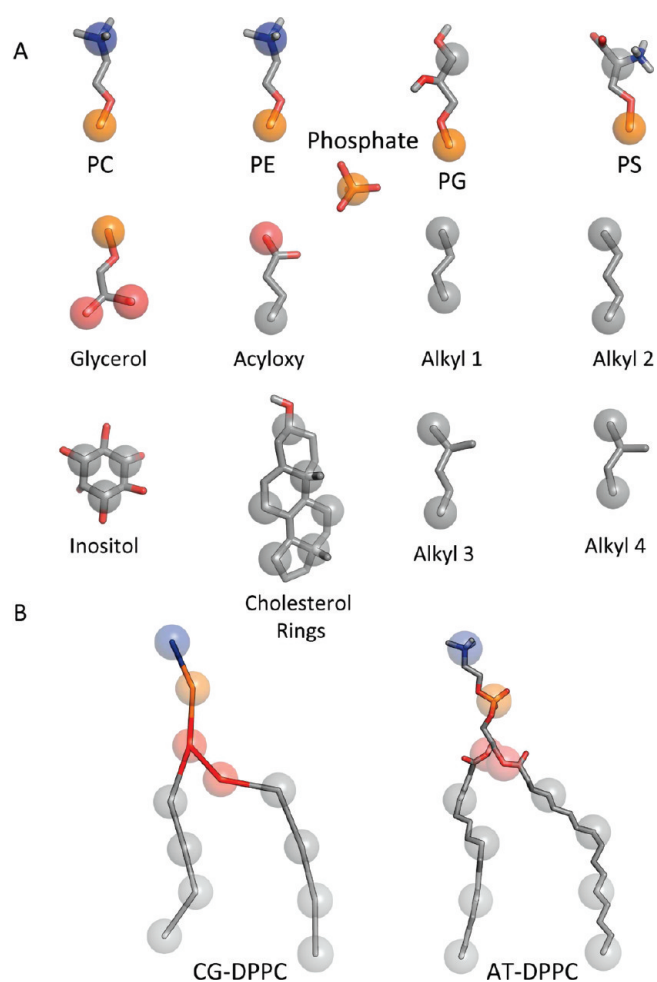


Figure 2. (A) Library of lipid fragments. Atomistic fragments used for the conversion of the lipid molecules. The atoms that are aligned with the CG particles are shown in spheres, with the remainder of the atoms shown as sticks. (B) An example of the lipid conversion is shown for DPPC.

memory Broyden–Fletcher–Goldfarb–Shanno (l-bfgs) energy minimization was required to reduce steric clashes between the headgroups of neighboring lipid molecules. Lipids currently available include POPC, POPE, POPG, POPS, DPPC, DHPC, DMPC, DMPG, DOPC, DSPE, BNG (β -nonyl glucoside), cholesterol, PIP₂, and PIP₃.

Final Steps. Following conversion of the protein and lipids the two energy minimized components were then combined and energy minimized further. In the case of protein converted using CG2AT-align, lipids within 1 Å of the protein were deleted to remove any unavoidable protein/lipid clashes. The system is then solvated, with any waters sitting within the hydrophobic core of the bilayer removed. Counterions are added to neutralize the system. The solvated system was finally energy minimized in preparation for the MD simulations.

Atomistic MD Simulations. Atomistic MD simulations were performed using Gromacs v4.5.3 with the GROMOS96 43a2 force field.³⁴ Simulations were performed using semi-isotropic pressure coupling with the Parrinello–Rahman barostat,³⁵ while the temperature of the lipid, protein, and solvent (water and counterions) was separately coupled to an external bath held at 323 K, using the Berendsen thermostat.³⁶ The water model used

was SPC.³⁷ The LINCS algorithm was used to constrain bond lengths.³⁸ Long-range electrostatic interactions beyond 10 Å were modeled using the particle mesh Ewald (PME) method.³⁹ A cutoff of 10 Å was used for van der Waals' interactions. Each converted system was first subjected to 1 ns of protein-restrained simulation, during which all heavy atoms of the protein were harmonically restrained with a force constant of 1000 kJ/mol/nm³. These restraints were then removed for 50 ns of production simulation with coordinates saved every 10 ps for analysis.

Pure Lipid Bilayer Simulations. In addition to studying the conversion of membrane protein complexes we also assessed this methodology with pure lipid bilayers, consisting of either POPC, POPE, DOPC, DPPC, or DMPC. Each bilayer was self-assembled by a 500 ns CG-MD simulation, before conversion to an atomistic systems, and used to start a 10 ns AT-MD simulation. The final snapshot of these simulations, along with the parameters used, can be found in Lipidbook (<http://lipidbook.bioch.ox.ac.uk/>).⁴⁰

RESULTS

Example: An Ion Channel Protein. The method is best illustrated by following in detail a specific example. For this we have selected the ASIC protein. This is an acid-sensing ion channel the structure of which has been determined at 1.9 Å resolution.⁴¹ It is trimeric, with a transmembrane (TM) domain containing six helices and an extensive extracellular domain. It is therefore a good example of a moderately complex membrane protein. It has been the subject of some simulation studies.⁴² The conversion process is illustrated in Figure 3. It can be seen that the CG-MD self-assembly process 'correctly' inserts the protein in a bilayer, i.e., with presumed TM domain in a bilayer spanning orientation and with the extracellular domain making few contacts with the lipids.

ASIC is of interest in that the trimer is asymmetric. This is reflected in the asymmetric, tilted orientation of the trimer relative to the bilayer. This is seen in the CG-MD simulation and increases during the 50 ns AT-MD simulation (Figure 4A and B). The orientation during the AT-MD simulation may be compared with the orientation predicted in the OPM database (<http://opm.phar.umich.edu/>), which treats the bilayer via an implicit bilayer method.¹³ Both the simulations and the implicit bilayer method suggest that the TM domain of ASIC is tilted (Figure 4). However the degree to which the protein changes over the course of the atomistic simulations, as the ectodomain interacts with the interfacial region of the bilayer. Thus for the three subunits, considering the TM segments only, the simulation-based tilt angles (relative to the bilayer normal, averaged over the last 10 ns of the two AT-MD simulations) are: subunit A, ~19°; subunit B, ~60°; and subunit C, ~48°. For OPM the corresponding tilt angles are: 2°, 25°, and 25°. In either case, the tilt is greater than that which would be assumed by simple 'manual' positioning of the ASIC protein in a lipid bilayer, guided by inspection transmembrane regions of the X-ray structure (Figure 4D). This difference in tilt angle is likely to be important in functional predictions of, e.g., the electrostatic potential surface around the protein where is embedded in the bilayer (see e.g. ref 42).

We can compare the lipid headgroup contacts predicted by CG-MD and those maintained in the AT-MD simulation of ASIC. It can be seen there is an excellent agreement between the lipids contacts suggested by the CG simulations and those

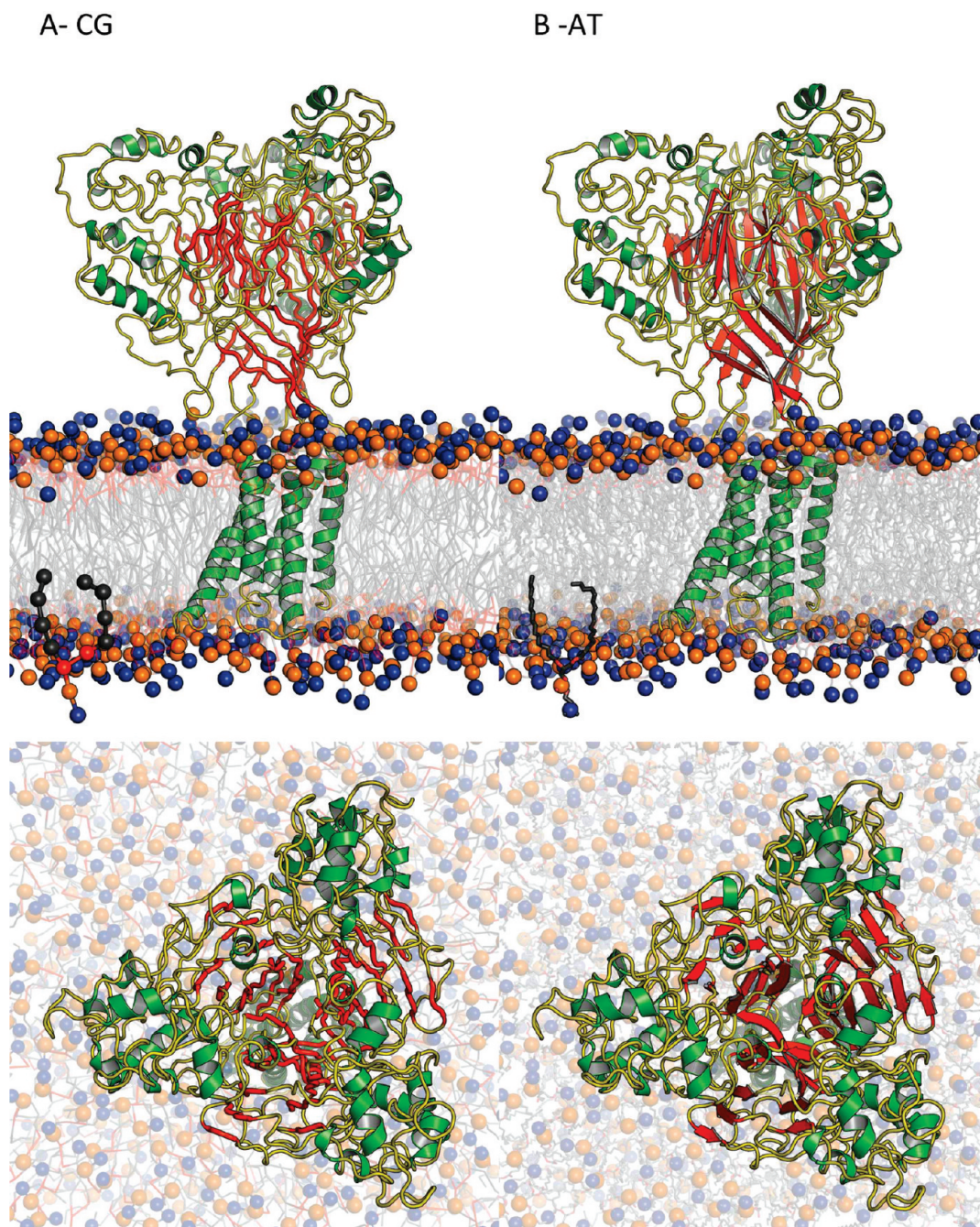


Figure 3. Visual comparison between CG and AT representations. A comparison between the (A, C) CG and (B, D) atomistic system of ASIC shown perpendicular to (top) and down (bottom) the bilayer normal.

preserved after CG-to-AT conversion and AT-MD simulation (Figure 5), regardless of whether the conversion used the ‘full’ or ‘aligned’ procedure for the protein (see above and Figure 1). This is reflected in correlation coefficients of between 0.79 to 0.81 for the lipid contacts seen in the three simulations.

Benchmark Systems. We employed a benchmark set of 10 membrane proteins against which to evaluate the CG2AT procedure and the behavior and lipid interactions of the proteins in the subsequent short AT-MD simulations (Figure 6). These 10 systems were selected to span a range of membrane proteins, simple and complex, with different overall architectures and differing patterns of interaction with lipids. Thus, there are

two relatively simple integral membrane proteins—for which most of the protein mass is α -helical and is located in the bilayer—namely LeuT and an aquaporin. There are three proteins with extensive extracellular (ELIC, ASIC) or equivalent (Cyt Ox) domains and three proteins with extensive intracellular domains (KcsA, SERCA and β_2 AdR/lysozyme). β_2 AdR/lysozyme is of especial interest as it is an artificial chimeric construct in which lysozyme has been inserted into an intracellular loop of a GPCR protein. The two other major classes of membranes proteins—outer membrane β -barrels and monotopic membrane proteins—are represented by OmpC and OSC, respectively.

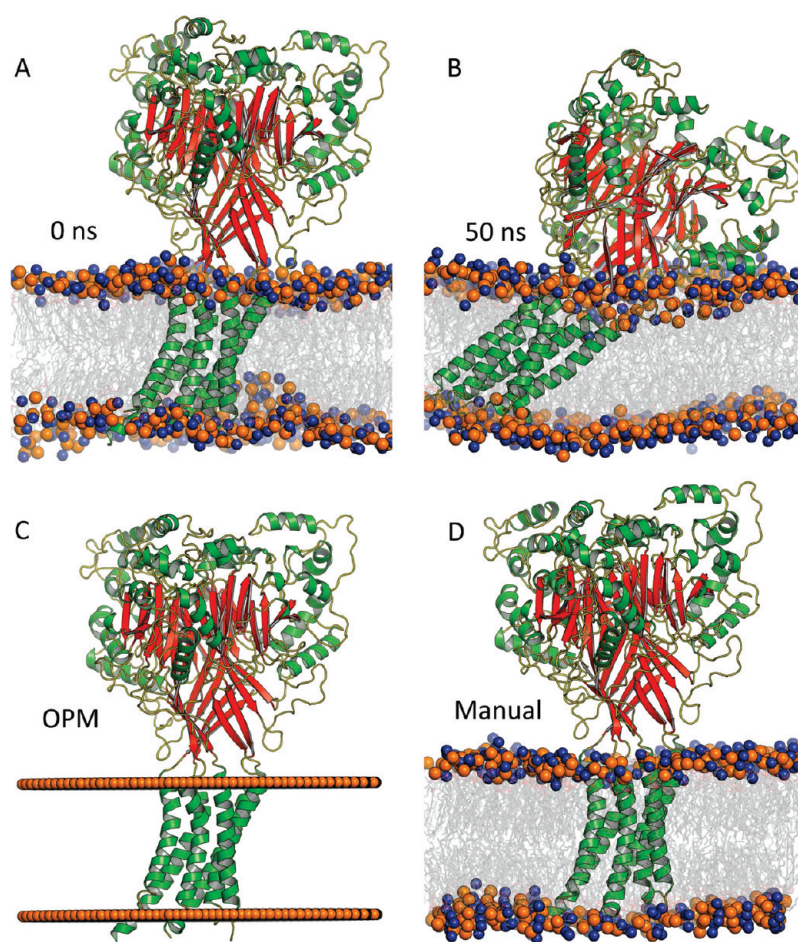


Figure 4. ASIC AT-MD simulations. Snapshots of ASIC in a lipid bilayer at the start (A) and end (B) of the AT-MD simulation from the CG2AT-align conversion. (C) ASIC oriented relative to the bilayer plane as predicted by OPM¹³ (<http://opm.phar.umich.edu/>). (D) ASIC oriented ‘manually’ in a bilayer based on helix locations in the X-ray structure and the sequence-based prediction of TM helices.

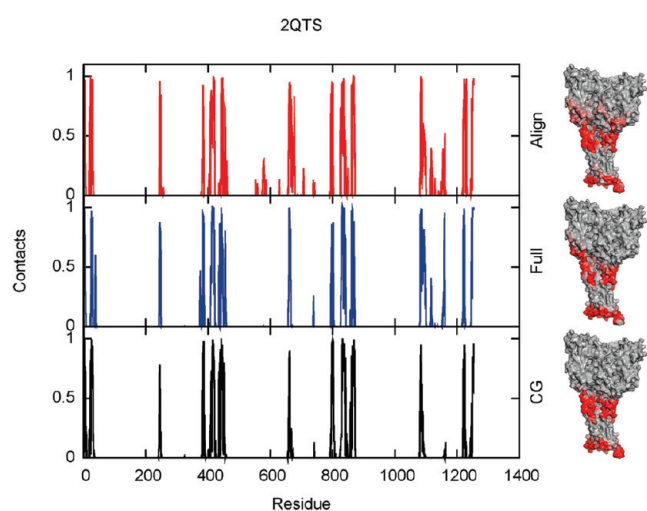


Figure 5. ASIC lipid/protein interactions. (A) CG and AT structures of ASIC color coded according to the frequency of lipid headgroup contacts averaged over the respective simulations; blue = no contacts made. (B) Lipid headgroup contact frequencies (as a fraction of simulation time) as a function of residue number for the CG-MD and the two AT-MD simulations of ASIC.

Behavior of the Proteins. One may compare the protein structures in terms of C α root-mean-square deviations (rmsds) before and after the CG2AT conversion. Unsurprisingly, the C α rmsd is small when the protein structure converted using Pulchra and Modeler, with an average value of 1.3 Å. For the CG2AT-align protocol the rmsd fit between the CG protein, and the overlaid X-ray structure is lower, with an average of 2.5 Å across all of the test set of proteins. This reflects a degree of (local) conformational change from the initial structures in the CG simulations, largely determined by the elastic network restraints used to model the protein tertiary structures in the CG method.

One may also monitor the degree of conformational drift over the course of the 50 ns AT-MD simulations. All simulations have a relatively low C α rmsd drift, especially for the transmembrane domains (Figure 7). Nevertheless, the overall degree of conformational drift is somewhat lower (by ca. 0.5 Å on average) for the simulations starting from the X-ray protein coordinates (i.e., from CG2AT-align) than those starting with remodeled protein coordinates (i.e., from CG2AT-full). An interesting outlier in terms of C α rmsd is the β_2 AdR/lysosyme chimeric protein. For this the rmsd is much lower if one considers only the β_2 -adrenoceptor domain within the chimeric construct.

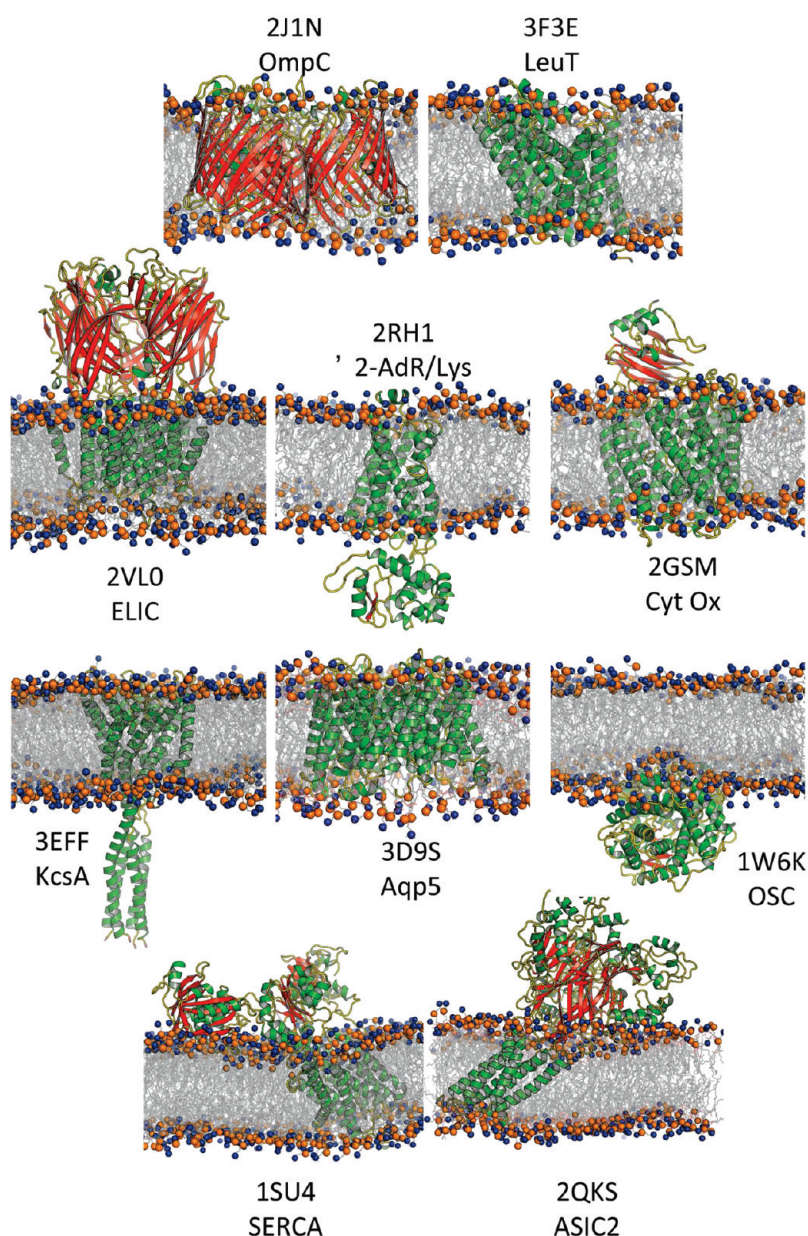


Figure 6. Benchmark systems. The 10 systems studied, from the AT-MD simulations. In each case the protein is shown by a ribbon, colored by secondary structure; green, α -helices; red, β -strands; and yellow, loops. The phosphorus (orange) and nitrogen (blue) atoms of the DPPC lipids are shown as spheres, to indicate the position of the bilayer.

Behavior of the Lipids. The average rmsd between CG and atomistic lipids was 1 Å, when aligning the atomic coordinates with their counterpart CG particle. One of the main benefits of the CGMD approach for simulating bilayer arrangement around a protein is it can allow for any local or global membrane deformation by the protein. The converted atomistic system (from CG2AT-full) retains the bilayer distortions of the CG simulation, allowing tight packing around the membrane protein. In contrast, in the CG2AT-align method on average 25 lipid molecules were deleted to remove any protein/lipid clashes. Although the deletion of the lipids removes close contacts between lipid and protein, these contacts are regained to a certain extent in the initial 1 ns equilibration step of the subsequent AT-MD simulation during which the protein is restrained. Such a lipid deletion step can be avoided by

reducing the protein flexibility during the CG-MD bilayer self-assembly step by, e.g., increasing the CG protein elastic network model⁴³ cutoff to ≥ 10 Å. This allows for better retention of the key protein/lipid interactions predicted by the CG-MD simulations while maintaining the X-ray structure of the protein.

We calculated the lipid tail order parameter (S_{CD} ; Figure 8) profiles from the AT-MD simulations both for pure lipid bilayers generated by CG2AT and for the two sets (align and full) of lipid/protein CG2AT systems. Those for the lipid only simulations closely resemble those from standard atomistic simulations of lipid bilayers thus suggesting that the fragment-based conversion methods from CG lipids preserves a stable fluid-phase lipid bilayer. The order parameter curves for proteins are very similar whichever CG2AT method is used for the protein and in

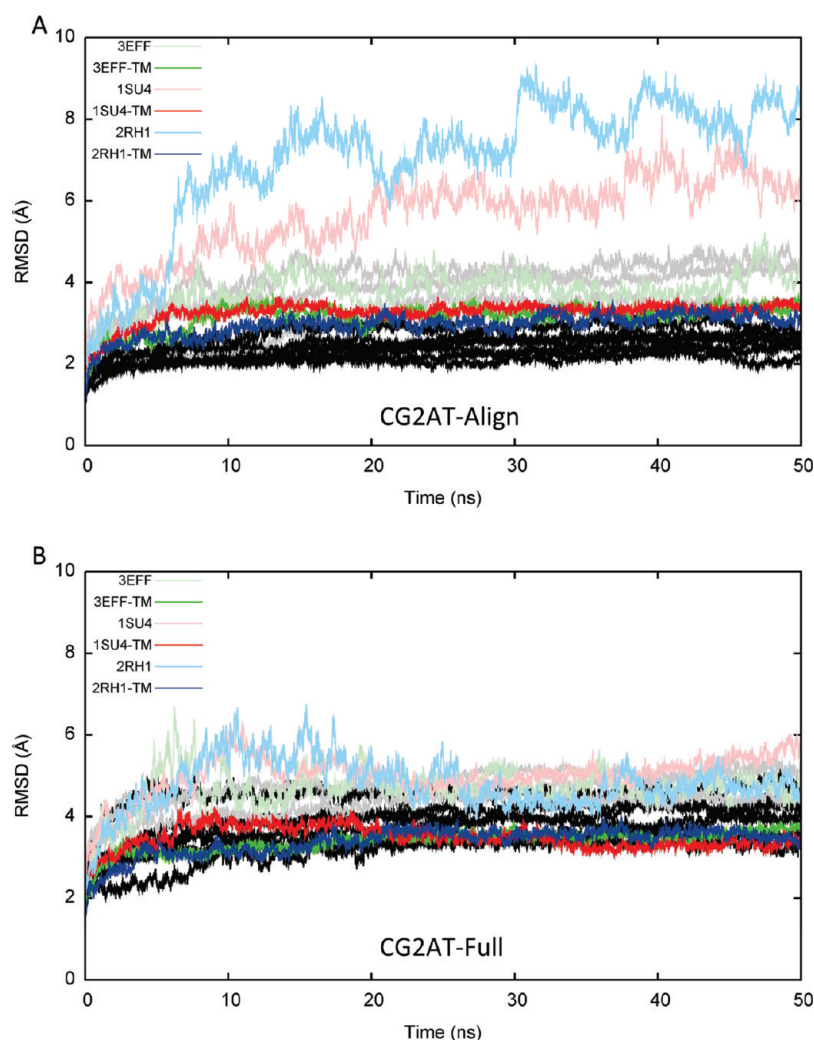


Figure 7. Protein conformational drift for the AT-MD simulations. The $C\alpha$ rmsd as a function of time the AT-MD simulations of the 10 systems are shown, using either the (A) CG2AT-align or (B) CG2AT-full method for protein conversion. The rmsds for the complete proteins are shown in gray, and those for the transmembrane domains only are in black. The β 2-adrenoceptor structure (2RH1) shows the greatest drift from the starting structure (light blue), when the entire engineered complex is considered. Nevertheless, if one considers only the GPCR component, the structure is very stable (red). The SERCA structure (1SU4; entire structure, pink; TM only, red) and full-length KcsA (3EFF; entire structure, light green; TM only, green) also shown to be relatively dynamic structures in the non-TM regions.

general suggest a degree of (local) increase in lipid tail order by the proteins, as might be anticipated.

Lipid/Protein Interactions. The lipid headgroups contacting the proteins were analyzed as a function of protein residue number in a similar fashion to that for ASIC (see above). It is evident that there are high-correlation coefficients between the contacts observed in the three simulations (Table 1). Therefore we may conclude that the CG2AT conversion procedure preserves the key lipid/protein contacts generated by self-assembly CG-MD and that these contacts remain, at least over the short (50 ns) duration of the AT-MD simulations.

We also analyzed the frequency distributions along the bilayer normal (z) of amino acid residue types that make contacts with the lipid molecules in the AT-MD simulations for at least 30% of the time (Supporting Information, Figure S2). These distributions show the same general patterns as seen in CG-MD simulations of a wide range of membrane proteins.¹⁶ Thus amphipathic aromatic residues (e.g., Trp and Tyr) are located in the lipid/water interfacial regions, as are Arg and Lys side chains (but

at higher somewhat $|z|$ values allowing interactions with lipid phosphate groups), whereas hydrophobic side chains are preferentially localized in the bilayer core. This further demonstrates that the CG2AT conversion has preserved the positions and the interactions of lipid-exposed amino acids on the surface of the TM domains of membrane proteins.

DISCUSSION

We have described a method for accurate and automatic conversion of a CG description of a complex membrane/protein system to atomistic detail, enabling subsequent AT-MD simulation. This method has been evaluated against a test set of 10 membrane proteins and shown to provide stable AT-MD simulation systems which in turn provide information on, e.g., protein/lipid interactions. Thus, we have provided a protocol for serial multiscale (as defined by Voth and colleagues)²³ simulation of membrane proteins. This combines efficient sampling of

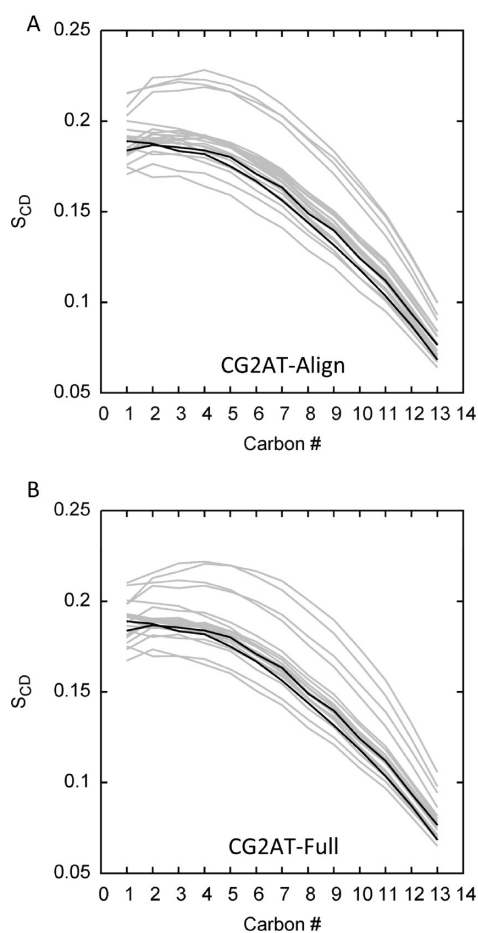


Figure 8. Lipid order parameters. The lipid order parameters (S_{CD}) from a 50 ns simulation of pure DPPC lipid bilayers (black) are compared to the order parameters derived from the systems containing protein (gray) that have been converted either through (A) structural alignment or (B) the full conversion method.

lipid/protein interactions in CG-MD simulations with refinement of these interactions on a shorter time scale by AT-MD.

Currently there are relatively few protocols available for preparing complex, mixed lipid systems around a membrane protein. OPM^{13,44} and related prediction methods indicate the regions of the protein that will interact with lipid, and implicit bilayer models^{45–47} can provide efficient sampling in a bilayer-like environment and allow use enhanced sampling methods, such as replica exchange.⁴⁸ However, particle-based approximations (either CG-MD as in the current study, or DPD as in, e.g., refs 49–52), are needed to provide models of more specific lipid/protein interactions, including local distortions of bilayer thickness and/or selective interactions with lipid headgroups. By combining such approaches with CG2AT, it is possible to initiate detailed atomistic simulations to refine the resultant models of such interactions.

We have described two variants on the CG2AT method, differing in whether or not changes in protein conformation from CG-MD are retained. The CG2AT-align method should be used where preserving the exact starting structure of the protein is important for the subsequent atomistic molecular simulations, e.g., if known ligand binding sites need to be accurately retained. Additionally, in order to retain the local protein/lipid

Table 1. Lipid/Protein Contacts: Correlations Between Simulations^a

protein	PDB id	correlation coefficient		
		CG vs AT-full	CG vs AT-align	AT-full vs AT-align
SERCA	1SU4	0.73	0.71	0.78
OSC	1W6K	0.66	0.53	0.67
Cyt Ox	2GSM	0.79	0.78	0.81
OmpC	2J1N	0.80	0.85	0.86
ASIC	2QTS	0.81	0.77	0.78
B ₂ AdR/Lys	2RH1	0.81	0.84	0.82
ELIC	2VL0	0.77	0.80	0.76
Aqp5	3D9S	0.80	0.79	0.76
KcsA	3EFF	0.70	0.54	0.59
LeuT	3F3E	0.77	0.78	0.81

^aThese proteins include calcium ATPase (SERCA) (1SU4),⁶¹ oxidosqualene cyclase (OSC) (PDB id: 1W6K),⁶² cytochrome c oxidase (2GSM),⁶³ outer membrane protease C (OmpC) (2J1N),⁶⁴ acid-sensing ion channel 2 (ASIC) (2QTS),⁴¹ β -2 adrenenergetic receptor with lysozyme (β 2-AdR/lys) (2RH1),⁶⁵ ELIC pentameric ion channel (2VL0),⁶⁶ aquaporin P5 (Aqp5) (3D9S),⁶⁷ KcsA potassium channel (3EFF),⁶⁸ and leucine transporter (LeuT) (3F3E).⁶⁹

contacts, the CG-MD simulations may be performed with tighter elastic network model⁴³ restraints imposed on the protein structure to prevent any major conformational changes and therefore avoid the need for lipid deletion upon conversion. In addition, the GROMACS *g_membed* protocol⁵³ has been included in more recent versions of the conversion process.

The major strength of our approach is that it enables detailed studies of lipid/protein interactions by providing an optimal orientation of a complex membrane protein in a bilayer for subsequent AT-MD simulations. This is illustrated here for the ion channel ASIC, enabling comparison of the multiscale simulation approach with other methods (e.g., ‘manual’ insertion or use of the OPM model) for prediction of how the protein might be arranged relative to a bilayer and indicating significant differences in the final result. In the current paper we have used a relatively simple lipid bilayer (DPPC), but recent studies have shown that a similar approach may be applied to more complex protein/lipid systems, e.g., models of Kir channels and their possible interactions with PIP₂.⁵⁴

On balance we think that our method for CG2AT conversion performs well for membrane protein systems, allowing a balance between utility and accuracy to be achieved in the context of a given biological problem. Thus the current method, based on fragment assembly, provides a possible alternative to other methods that have used simulated annealing⁵⁵ or force matching.⁵⁶ It also appears to be a less time-intensive mechanism for conversion that that used previously in our laboratory.⁵⁷ However, the current approach does not balance the energetics of both levels of granularity. An alternative approach would be to use mixed resolution methods which in principle could achieve detailed balance.^{58,59} However, at present such methods are likely to be difficult to applying to complex multicomponent membrane systems.

It should also be noted that we do not attempt to convert the water particles from the CG simulation, as these are somewhat

simplified in MARTINI and related CG models. However, it should be possible to include such a conversion if appropriate, e.g., if polarizable CG water models⁶⁰ are used.

As the number of membrane protein structures determined by X-ray crystallography and other high-resolution methods expands,³ there is an increasing need for multiscale simulations of membrane proteins. Building upon our earlier database of CG simulations (<http://sbcb.bioch.ox.ac.uk/cgdb/>),^{16,26} it should now be possible to provide multiscale simulations of all membrane proteins as their structures are determined. This in turn will enable structural bioinformatics studies, such as data mining of membrane protein/lipid interactions.

■ ASSOCIATED CONTENT

S Supporting Information. The CG particle to atomistic mapping for the protein sidechains for (A) our in house CG protein parameters and (B) Martini. Distribution along the bilayer normal of amino acids making contact with the bilayer over the course of the two sets of twenty 10 ns atomistic simulations of the membrane protein complexes. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: mark.sansom@bioch.ox.ac.uk

■ ACKNOWLEDGMENT

We thank members of MSPS laboratory for helpful discussions. Research in MSPS laboratory is supported by the Wellcome Trust and BBSRC.

■ REFERENCES

- (1) Wallin, E.; von Heijne, G. Genome-wide analysis of integral membrane proteins from eubacterial, archean, and eukaryotic organisms. *Protein Sci.* **1998**, *7*, 1029–1038.
- (2) Terstappen, G. C.; Reggiani, A. In silico research in drug discovery. *Trends Pharmacol. Sci.* **2001**, *22*, 23–26.
- (3) White, S. H. Biophysical dissection of membrane proteins. *Nature* **2009**, *459*, 344–6.
- (4) Gonen, T.; Cheng, Y.; Sliz, P.; Hiroaki, Y.; Fujiyoshi, Y.; Harrison, S. C.; Walz, T. Lipid–protein interactions in double-layered two-dimensional AQP0 crystals. *Nature* **2005**, *438*, 633–638.
- (5) Hite, R. K.; Li, Z. L.; Walz, T. Principles of membrane protein interactions with annular lipids deduced from aquaporin-0 2D crystals. *EMBO J.* **2009**, *29*, 1652–1658.
- (6) Powl, A. M.; Wright, J. N.; East, J. M.; Lee, A. G. Identification of the hydrophobic thickness of a membrane protein using fluorescence spectroscopy: Studies with the mechanosensitive channel MscL. *Biochemistry* **2005**, *44*, 5713–5721.
- (7) Anbazhagan, V.; Qu, J.; Kleinschmidt, J. H.; Marsh, D. Incorporation of outer membrane protein OmpG in lipid membranes: Protein-lipid interactions and beta-barrel orientation. *Biochemistry* **2008**, *47*, 6189–6198.
- (8) Gold, V. A. M.; Robson, A.; Bao, H.; Romantsov, T.; Duong, F.; Collinson, I. The action of cardiolipin on the bacterial translocon. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 10044–10049.
- (9) Lindahl, E.; Sansom, M. S. P. Membrane proteins: molecular dynamics simulations. *Curr. Opin. Struct. Biol.* **2008**, *18*, 425–431.
- (10) Niemela, P. S.; Miettinen, M. S.; Monticelli, L.; Hammaren, H.; Bjelkmar, P.; Murtola, T.; E, L.; Vattulainen, I. Membrane proteins diffuse as dynamic complexes with lipids. *J. Am. Chem. Soc.* **2010**, *132*, 7574–7575.
- (11) Faraldo-Gómez, J. D.; Smith, G. R.; Sansom, M. S. P. Setup and optimization of membrane protein simulations. *Eur. Biophys. J.* **2002**, *31*, 217–227.
- (12) Killian, J. A.; von Heijne, G. How proteins adapt to a membrane-water interface. *Trends Biochem. Sci.* **2000**, *25*, 429–434.
- (13) Lomize, M. A.; Lomize, A. L.; Pogozheva, I. D.; Mosberg, H. I. OPM: Orientations of proteins in membranes database. *Bioinformatics* **2006**, *22*, 623–625.
- (14) Voth, G. A. *Coarse-Graining of Condensed Phase and Biomolecular Systems*; CRC Press: Boca Raton, FL, 2008.
- (15) Periole, X.; Huber, T.; Marrink, S. J.; Sakmar, T. P. G protein-coupled receptors self-assemble in dynamics simulations of model bilayers. *J. Am. Chem. Soc.* **2007**, *129*, 10126–10132.
- (16) Scott, K. A.; Bond, P. J.; Ivetac, A.; Chetwynd, A. P.; Khalid, S.; Sansom, M. S. P.; Coarse-grained, M. D. simulations of membrane protein-bilayer self-assembly. *Structure* **2008**, *16*, 621–630.
- (17) Vorobyov, I.; Li, L.; Allen, T. W. Assessing atomistic and coarse-grained force fields for protein-lipid interactions: the formidable challenge of an ionizable side chain in a membrane. *J. Phys. Chem. B* **2008**, *112*, 9588–9602.
- (18) Bond, P. J.; Wee, C. L.; Sansom, M. S. P. Coarse-grained molecular dynamics simulations of the energetics of helix insertion into a lipid bilayer. *Biochemistry* **2008**, *47*, 11321–11331.
- (19) Shi, Q.; Izvekov, S.; Voth, G. A. Mixed atomistic and coarse-grained molecular dynamics: simulation of a membrane bound ion channel. *J. Phys. Chem. B* **2006**, *110*, 15045–15048.
- (20) Lu, L. Y.; Izvekov, S.; Das, A.; Andersen, H. C.; Voth, G. A. Efficient, regularized, and scalable algorithms for multiscale coarse-graining. *J. Chem. Theory Comp.* **2010**, *6*, 954–965.
- (21) Orsi, M.; Sanderson, W. E.; Essex, J. W. Permeability of small molecules through a lipid bilayer: a multiscale simulation study. *J. Phys. Chem. B* **2009**, *113*, 12019–12029.
- (22) Orsi, M.; Essex, J. W. Permeability of drugs and hormones through a lipid bilayer: insights from dual-resolution molecular dynamics. *Soft Matter* **2010**, *6*, 3797–3808.
- (23) Ayton, G. A.; Noid, W. G.; Voth, G. A. Multiscale modeling of biomolecular systems: in serial and in parallel. *Curr. Opin. Struct. Biol.* **2007**, *17*, 192–198.
- (24) Thogersen, L.; Schiott, B.; Vosegaard, T.; Nielsen, N. C.; Tajkhorshid, E. Peptide aggregation and pore formation in a lipid bilayer - a combined coarse grained and all atom molecular dynamics study. *Biophys. J.* **2008**, *95*, 4337–4347.
- (25) Rzepiela, A. J.; Schafer, L. V.; Goga, N.; Risselada, H. J.; De Vries, A. H.; Marrink, S. J. Reconstruction of atomistic details from coarse-grained structures. *J. Comput. Chem.* **2010**, *31*, 1333–1343.
- (26) Chetwynd, A. P.; Scott, K. A.; Mokrab, Y.; Sansom, M. S. P. CGDB: a database of membrane protein/lipid interactions by coarse-grained molecular dynamics simulations. *Mol. Membr. Biol.* **2008**, *25*, 662–669.
- (27) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. GROMACS: fast, flexible, and free. *J. Comput. Chem.* **2005**, *26*, 1701–1718.
- (28) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. The MARTINI forcefield: coarse grained model for biomolecular simulations. *J. Phys. Chem. B* **2007**, *111*, 7812–7824.
- (29) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S. J. The MARTINI coarse grained force field: extension to proteins. *J. Chem. Theor. Comp.* **2008**, *4*, 819–834.
- (30) Sali, A.; Blundell, T. L. Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Biol.* **1993**, *234*, 779–815.
- (31) Rotkiewicz, P.; Skolnick, J. Fast procedure for reconstruction of full-atom protein models from reduced representations. *J. Comput. Chem.* **2008**, *29*, 1460–5.
- (32) Laskowski, R. A.; Macarthur, M. W.; Moss, D. S.; Thornton, J. M. Procheck - a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **1993**, *26*, 283–291.

- (33) Cuthbertson, J. M.; Doyle, D. A.; Sansom, M. S. P. Transmembrane helix prediction: a comparative evaluation and analysis. *Prot. Eng. Des. Sel.* **2005**, *18*, 295–308.
- (34) Scott, W. R. P.; Hunenberger, P. H.; Tironi, I. G.; Mark, A. E.; Billeter, S. R.; Fennen, J.; Torda, A. E.; Huber, T.; Kruger, P.; van Gunsteren, W. F. The GROMOS biomolecular simulation program package. *J. Phys. Chem. A* **1999**, *103*, 3596–3607.
- (35) Parrinello, M.; Rahman, A. Polymorphic transitions in single-crystals - a new molecular-dynamics method. *J. Appl. Phys.* **1981**, *52*, 7182–7190.
- (36) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
- (37) Hermans, J.; Berendsen, H. J. C.; van Gunsteren, W. F.; Postma, J. P. M. A consistent empirical potential for water-protein interactions. *Biopolymers* **1984**, *23*, 1513–1518.
- (38) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **1997**, *18*, 1463–1472.
- (39) 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*, 10089–10092.
- (40) Domański, J.; Stansfeld, P. J.; Sansom, M. S. P.; Beckstein, O. Lipidbook - a public repository for force field parameters used in membrane simulations. *J. Membr. Biol.* **2010**, *236*, 255–258.
- (41) Jasti, J.; Furukawa, H.; Gonzales, E. B.; Gouaux, E. Structure of acid-sensing ion channel 1 at 1.9 Å resolution and low pH. *Nature* **2007**, *449*, 316–23.
- (42) Shaikh, S. A.; Tajkhorshid, E. Potential cation and H^+ binding sites in acid sensing ion channel-1. *Biophys. J.* **2008**, *95*, 5153–5164.
- (43) Atilgan, A. R.; Durell, S. R.; Jernigan, R. L.; Demirel, M. C.; Keskin, O.; Bahar, I. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.* **2001**, *80*, 505–515.
- (44) Lomize, A. L.; Pogozheva, I. D.; Lomize, M. A.; Mosberg, H. I. Positioning of proteins in membranes: A computational approach. *Protein Sci.* **2006**, *15*, 1318–1333.
- (45) Ulmschneider, M. B.; Sansom, M. S. P.; Di Nola, A. Properties of integral membrane protein structures: derivation of an implicit membrane potential. *Proteins: Struct., Funct., Bioinf.* **2005**, *59*, 252–265.
- (46) Ulmschneider, M. B.; Ulmschneider, J. P.; Sansom, M. S. P.; Di Nola, A. A generalized Born implicit membrane representation based on experimental insertion free energies. *Biophys. J.* **2007**, *92*, 2338–2349.
- (47) Tanizaki, S.; Feig, M. Molecular dynamics simulations of large integral membrane proteins with an implicit membrane model. *J. Phys. Chem. B* **2006**, *110*, 548–556.
- (48) Sayadi, M.; Tanizaki, S.; Feig, M. Effect of membrane thickness on conformational sampling of phospholamban from computer simulations. *Biophys. J.* **2010**, *98*, 805–814.
- (49) de Meyer, F. J. M.; Venturoli, M.; Smit, B. Molecular simulations of lipid-mediated protein-protein interactions. *Biophys. J.* **2008**, *95*, 1851–1865.
- (50) de Meyer, F. J. M.; Rodgers, J. M.; Willems, T. F.; Smit, B. Molecular simulation of the effect of cholesterol on lipid-mediated protein-protein interactions. *Biophys. J.* **2010**, *99*, 3629–3638.
- (51) Schmidt, U.; Guigas, G.; Weiss, M. Cluster formation of transmembrane proteins due to hydrophobic matching. *Phys. Rev. Lett.* **2008**, *101*, 128104.
- (52) Schmidt, U.; Weiss, M. Hydrophobic mismatch-induced clustering as a primer for protein sorting in the secretory pathway. *Biophys. Chem.* **2010**, *151*, 34–38.
- (53) Wolf, M. G.; Hoefling, M.; Aponte-Santamaría, C.; Grubmüller, H.; Groenhof, G. g_mbed : Efficient insertion of a membrane protein into an equilibrated lipid bilayer with minimal perturbation. *J. Comput. Chem.* **2010**, *31*, 2169–2174.
- (54) Stansfeld, P. J.; Hopkinson, R. J.; Ashcroft, F. M.; Sansom, M. S. P. The PIP_2 binding site in Kir channels: definition by multi-scale biomolecular simulations. *Biochemistry* **2009**, *48*, 10926–10933.
- (55) Rzepiela, A. J.; Schäfer, L. V.; Goga, N.; Risselada, H. J.; de Vries, A. H.; Marrink, S. J. Reconstruction of atomistic details from coarse grained structures. *J. Comput. Chem.* **2010**, *31*, 1333–1343.
- (56) Izvekov, S.; Voth, G. A. A multiscale coarse-graining method for biomolecular systems. *J. Phys. Chem. B* **2005**, *109*, 2469–2473.
- (57) Carpenter, T.; Bond, P. J.; Khalid, S.; Sansom, M. S. P. Self-assembly of a simple membrane protein: coarse-grained molecular dynamics simulations of the influenza M2 channel. *Biophys. J.* **2008**, *95*, 3790–3801.
- (58) Liu, P.; Voth, G. A. Smart resolution exchange: an efficient algorithm for exploring complex energy landscapes. *J. Chem. Phys.* **2007**, *126* (04S106), 1–6.
- (59) Liu, P.; Shi, Q.; Lyman, E.; Voth, G. A. Reconstructing atomistic detail from coarse-grained models with resolution exchange. *J. Chem. Phys.* **2008**, *129* (114103), 1–8.
- (60) Yesylevskyy, S. O.; Schäfer, L. V.; Sengupta, D.; Marrink, S. J. Polarizable water model for the coarse-grained MARTINI force field. *PLoS Comp. Biol.* **2010**, *6*, e1000810.
- (61) Toyoshima, C.; Nakasako, M.; Nomura, H.; Ogawa, H. Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 Å resolution. *Nature* **2000**, *405*, 647–655.
- (62) Thoma, R.; Schulz-Gasch, T.; D'Arcy, B.; Benz, J.; Aebi, J.; Dehmlow, H.; Hennig, M.; Stihle, M.; Ruf, A. Insight into steroid scaffold formation from the structure of human oxidosqualene cyclase. *Nature* **2004**, *432*, 118–122.
- (63) Qin, L.; Hiser, C.; Mulichak, A.; Garavito, R. M.; Ferguson-Miller, S. Identification of conserved lipid/detergent-binding sites in a high-resolution structure of the membrane protein cytochrome c oxidase. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 16117–22.
- (64) Basle, A.; Rummel, G.; Storici, P.; Rosenbusch, J. P.; Schirmer, T. Crystal structure of osmoporin OmpC from *E. coli* at 2.0 Å. *J. Mol. Biol.* **2006**, *362*, 933–942.
- (65) Cherezov, V.; Rosenbaum, D. M.; Hanson, M. A.; Rasmussen, S. G.; Thian, F. S.; Kobilka, T. S.; Choi, H. J.; Kuhn, P.; Weis, W. I.; Kobilka, B. K.; Stevens, R. C. High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. *Science* **2007**, *318*, 1258–65.
- (66) Hilf, R. J. C.; Dutzler, R. X-ray structure of a prokaryotic pentameric ligand-gated ion channel. *Nature* **2008**, *452*, 375–379.
- (67) Horsefield, R.; Norden, K.; Fellert, M.; Backmark, A.; Tomroth-Horsefield, S.; Terwisscha van Scheltinga, A. C.; Kvassman, J.; Kjellbom, P.; Johanson, U.; Neutze, R. High-resolution x-ray structure of human aquaporin 5. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 13327–32.
- (68) Uysal, S.; Vasquez, V.; Tereshko, V.; Esaki, K.; Fellouse, F. A.; Sidhu, S. S.; Koide, S.; Perozo, E.; Kossiakoff, A. Crystal structure of full-length KcsA in its closed conformation. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 6644–9.
- (69) Singh, S. K.; Piscitelli, C. L.; Yamashita, A.; Gouaux, E. A competitive inhibitor traps LeuT in an open-to-out conformation. *Science* **2008**, *322*, 1655–61.

GRIFFIN: A Versatile Methodology for Optimization of Protein–Lipid Interfaces for Membrane Protein Simulations

René Staritzbichler,[†] Claudio Anselmi,[‡] Lucy R. Forrest,^{*,†} and José D. Faraldo-Gómez^{*,†,§}

[†]Computational Structural Biology Group and [‡]Theoretical Molecular Biophysics Group, Max Planck Institute of Biophysics, Frankfurt am Main, Germany

[§]Cluster of Excellence Macromolecular Complexes, Frankfurt am Main, Germany

ABSTRACT: As new atomic structures of membrane proteins are resolved, they reveal increasingly complex transmembrane topologies and highly irregular surfaces with crevices and pores. In many cases, specific interactions formed with the lipid membrane are functionally crucial, as is the overall lipid composition. Compounded with increasing protein size, these characteristics pose a challenge for the construction of simulation models of membrane proteins in lipid environments; clearly, that these models are sufficiently realistic bears upon the reliability of simulation-based studies of these systems. Here, we introduce GRIFFIN (GRID-based Force Field INput), which uses a versatile framework to automate and improve a widely used membrane-embedding protocol. Initially, GRIFFIN carves out lipid and water molecules from a volume equivalent to that of the protein, to conserve the system density. In the subsequent optimization phase GRIFFIN adds an implicit grid-based protein force field to a molecular dynamics simulation of the precarved membrane. In this force field, atoms inside the implicit protein volume experience an outward force that will expel them from that volume, whereas those outside are subject to electrostatic and van der Waals interactions with the implicit protein. At each step of the simulation, these forces are updated by GRIFFIN and combined with the intermolecular forces of the explicit lipid–water system. This procedure enables the construction of realistic and reproducible starting configurations of the protein–membrane interface within a reasonable time frame and with minimal intervention. GRIFFIN is a stand-alone tool designed to work alongside any existing molecular dynamics package, such as NAMD or GROMACS.

INTRODUCTION

Membrane proteins constitute around a third of all proteins encoded in a typical genome,^{1–3} and yet the microscopic mechanisms of their functions are only just beginning to be described. Major advances in such understanding have been made through the elucidation of the three-dimensional atomic structure of some of these proteins by, e.g., X-ray crystallography. Indeed, the exponential increase in the number of structures being reported⁴ highlights the very significant progress made in this area. However, crystallographic structures capture a single state of what is typically a dynamic conformational equilibrium, intrinsic to the functional mechanism of the protein. A variety of structure-based computational approaches focus on these dynamic properties, in order to complement the experimental data. Prominent among these approaches is molecular dynamics (MD) simulation, for its ability to provide insights at atomic resolution, and its robust and versatile theoretical framework.

A surprising feature of many of the newly discovered membrane protein structures is their complex transmembrane topology, and the highly irregular interfaces they appear to form with the surrounding lipid bilayer. Indeed, for some proteins, such as the voltage-gated and mechanosensitive channels, or osmoregulatory transporters, regulatory mechanisms dependent on lipid composition are likely to be conveyed precisely by this protein–lipid interface.^{5–8} As the structure of the protein–membrane interface is not known experimentally, such complexity poses a serious challenge for the construction of molecular-simulation models. At the same time, whether this interface is realistically

modeled will clearly influence the ability of simulation-based studies to derive plausible mechanistic hypotheses.

A number of methodologies have been proposed for the construction of lipid–protein complexes for simulation.^{9–14} These adopt one of two general strategies, namely either to assemble a lipid bilayer around the protein *de novo*, or to adapt existing and well-optimized membrane models to the protein structure. The second strategy is an appealing and efficient option, because it builds upon much effort and considerable success in the construction of realistic membrane models for MD simulations, for a range of different lipids or mixtures thereof. Atomic coordinates and force fields for such systems are publicly available.

In one of these adaptive strategies, the protein and hydrated lipid bilayer systems are first superimposed in the same coordinate space, and some overlapping lipids and waters are removed, as necessary to preserve the system density.^{12,14} However, the cavity thus carved in the bilayer system does not normally match the shape of the protein, because of the very irregular conformations adopted by lipid hydrocarbon tails. To resolve this problem, one of the authors helped develop a methodology¹² in which the surface of the protein is used in the course of a molecular dynamics simulation to define additional forces that expel these tails from the protein volume. In this way, lipid and solvent atoms become adapted perfectly to the shape of the protein, and there is minimal perturbation of the existing, preoptimized membrane model.

Received: October 7, 2010

Published: March 29, 2011

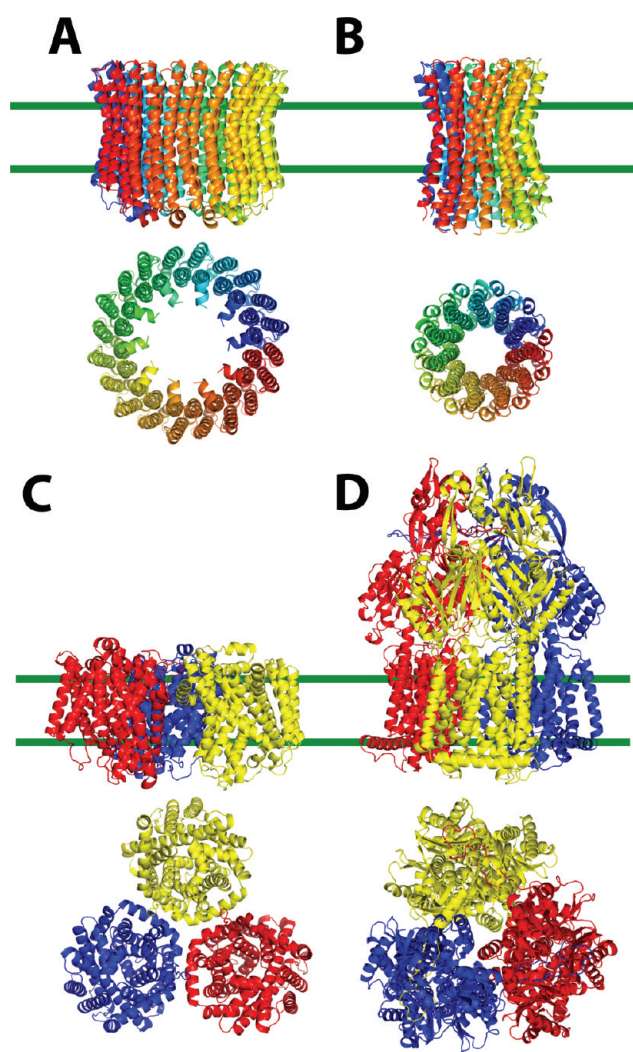


Figure 1. Representative membrane proteins with complex protein–lipid interfaces, used here as test cases for GRIFFIN. Each structure is displayed in cartoon format with different colors for each protein subunit, and viewed either from the plane of the membrane (upper panels) or normal to the membrane (lower panels). The approximate limits of the hydrophobic core of a hypothetical lipid membrane are also indicated (green lines). (A) The K_{10} rotor ring of the V-type Na^+ -ATPase from *Enterococcus hirae*. (B) The c_{11} rotor ring of the F-type Na^+ -ATPase from *Ilyobacter tartaricus*. (C) The carnitine/ γ -butyrobetaine antiporter CaiT from *Proteus mirabilis*. (D) The multi-drug efflux pump AcrB from *E. coli*.

Here, we present an improved and more general version of this surface-based approach, based upon a newly developed tool named GRIFFIN (GRId-based Force Field INput). GRIFFIN provides several significant advantages over the earlier implementation. First, forces acting on lipid atoms due to Coulomb and van der Waals interactions with the protein volume are now included along with the repulsive surface-guided forces; this results in chemical specificity at the protein–lipid interface, in addition to shape complementarity. Second, new features are included to selectively guide lipids or other molecules into or out of user-specified regions of the system, and also to prevent trapping of lipid tails in regions of high irregularity. These features thus provide a means to handle the kind of complex protein topologies increasingly uncovered by structural studies.

Third, GRIFFIN is designed to be a stand-alone, fully integrated application, compatible with any MD engine that provides a suitable interface, such as NAMD,¹⁵ GROMACS,¹⁶ or CHARMM.¹⁷ Last, its object-oriented grid-based algorithm, and the resulting computational efficiency, extends the applicability of the method to very large membrane protein complexes.

We demonstrate the effectiveness of this improved methodology with four examples, namely, the secondary transporter CaiT,¹⁸ two membrane rotors from the ATPase family,^{19,20} and the AcrB multidrug efflux pump²¹ (Figure 1). In all cases, irregular features of the protein–lipid interface make other approaches either unsuitable or impractical. These features include a central pore in the ATPase rings, and intersubunit channels and water-filled crevices in the trimeric CaiT and AcrB; AcrB is also one of the largest and most complex membrane proteins structures resolved to date. We will show how GRIFFIN enables the construction of physically realistic starting configurations of the lipid–protein interface for membrane protein simulations, following a reproducible procedure and within an affordable time frame, even for these very challenging cases.

METHODOLOGY

The overall protocol is described in the scheme in Figure 2. Initially, a cavity is carved into a hydrated lipid bilayer system, in which the membrane protein will ultimately be embedded. The protein structure, or rather, its surface, is used to estimate the number of lipid and water molecules to be carved out. In most cases, however, the resulting protein–lipid interface will be unsuitable for energy minimization, let alone simulation. The same surface is therefore employed to define a grid-based force field designed to expel the remaining lipid and/or water atoms from that volume. This force field also includes physical interactions (electrostatic and van der Waals), but only in the region of the grid outside the protein volume. During a subsequent molecular dynamics simulation of the precarved membrane system, the implicit protein force field is added to the standard interaction forces, thus optimizing the protein–lipid and protein–water interfaces. Geometric objects may be used as additional, lipid and/or water specific exclusion volumes, during the carving and/or the simulation.

Initial Carving of the Lipid Membrane. GRIFFIN provides an integrated tool for constructing the surface of the protein and for the initial carving of the membrane. The surface is constructed using a rolling-sphere method,²² with a probe of adjustable radius, typically set to 1.4 Å. Atomic radii are derived from the force field to be used in the simulation, e.g., CHARMM27.²³ The protein surface is used to estimate the volume occupied by the protein within each leaflet of the lipid membrane, as well in the hydration layer. The boundaries of these regions may be user-defined or determined by the program. Based on this volume estimation and the lipid density, the number of lipid molecules to be deleted from each leaflet is determined. These are selected from a list of all lipids that overlap with the protein volume, ranked according to the number of overlapping atoms and their distance to the protein surface. All overlapping water molecules are also deleted. Specific values for these parameters may be also provided by the user, overriding those calculated by the program.

Additional Molecule Type Specific Exclusion Volumes. GRIFFIN offers the possibility of using geometric objects (spheres, rectangular cuboids, and cones) as additional exclusion

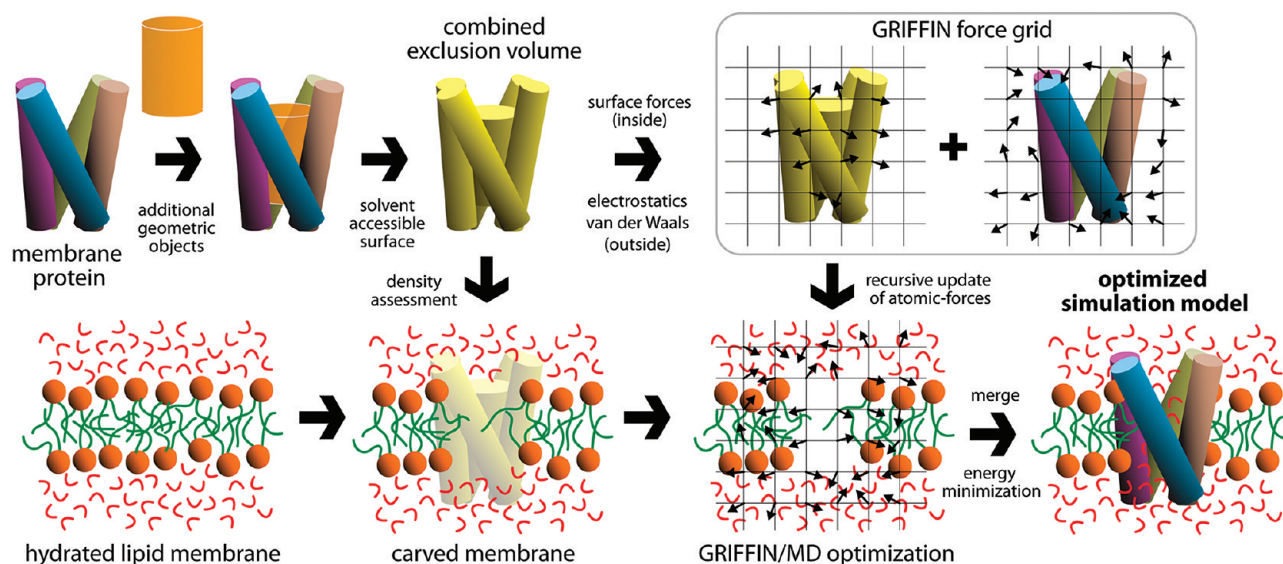


Figure 2. Schematic of the protocol for optimization of a protein–membrane interface using GRIFFIN. In the first step, a hypothetical membrane protein with a central pore (top left) is overlaid with a cylindrical geometric object (orange). The combined volume (yellow) of the protein and object is carved from an explicit system (bottom left) containing water (red) and lipid molecules (hydrocarbon chains are colored green and polar head groups are shown as orange spheres). After carving, some lipid and water atoms typically overlap the protein volume. The combined exclusion volume, and the actual protein structure, are also used to construct a force grid (top right) consisting of surface-directed expelling forces inside the volume and physical interaction forces outside. This force grid is overlaid on the molecular system to compute, at each time step of a molecular dynamics simulation, the actual forces on individual atoms; these are then added to the interaction energies between explicit atoms computed by the molecular dynamics package. In the last step, the GRIFFIN-optimized hydrated membrane is merged with the atomistic model of the membrane protein, and the entire system is energy-minimized to obtain the initial configuration for simulation (bottom right).

regions in the membrane carving, as well as in the optimization stage (see below). These objects are integrated in the volume estimation described above, and may be specific to a user-defined molecule type, e.g., to exclude lipid from a channel pore.

Calculation of the Implicit Protein Force Field. A second tool within GRIFFIN enables the calculation of the implicit force field for use during the subsequent molecular dynamics optimization stage. This force field is stored on a three-dimensional grid (Figure 3) of user-defined grid-point spacing (with a default value of 0.5 Å). Within the protein volume (which may be modified to include the optional exclusion regions mentioned above), each grid point stores a force (of magnitude 1 kcal/mol/Å) directed toward the nearest point on the protein surface (NPS; Figure 4A). These so-called surface forces will be applied to lipid and water atoms inside the volume during the optimization. Grid points outside of the protein surface map the Coulombic and van der Waals interaction forces between the protein and a probe particle of $q = 1e$, $\epsilon = 1$ kcal/mol, and $\sigma = 1$ Å (Figure 3). User-defined cutoff distances may be specified for each of these interactions (default values are 18, 12, and 8 Å, respectively). The calculation of the force grid is carried out only once during the protocol; however, as it may be time-consuming, this calculation may be distributed over an adjustable number of processors.

Calculation of GRIFFIN Forces during Molecular Dynamics Simulations. GRIFFIN supplies a molecular dynamics simulation with atomic forces derived from the precomputed grid. These are calculated at every simulation step, based on the atomic coordinates of the explicit system (Figure 3F). To map the grid-point forces onto the actual atoms, GRIFFIN uses a trilinear interpolation method. By default, the physical forces are scaled by the actual charge and van der Waals parameters of each atom (a geometric-mean combination rule is used for both

σ and ϵ). Surface forces may also be scaled by a constant; in practice this constant is typically increased in a stepwise manner, as the surface forces ultimately become balanced out by the external pressure (see Results). A user-specified constant factor may also be used to further scale each of the physical interaction forces.

To improve the performance and stability of the algorithm, no surface forces are applied to hydrogen atoms by default. By contrast, all atoms may experience physical forces if outside the protein volume. However, for greater efficiency, physical interactions are computed only for atoms neighboring the protein; a list of such atoms is updated every few steps during the simulation (typically 10), and includes all atoms for which the GRIFFIN force is nonzero at the update step. Lastly, it should be noted that by construction no reaction forces on the protein are considered; it is thus advisable that GRIFFIN is employed alongside stochastic dynamics, with tight temperature and pressure coupling.

A feature specific to lipid molecules is the possibility of redirection of the surface forces. This feature is important when atoms within the same lipid chain are driven in opposite directions (see Results). Redirection is applied when the surface force acting on a given atom is directed away from the overall geometric center of the lipid molecule to which it belongs (Figure 4). An adjustable threshold defines the maximum angle by which the surface force may diverge; this is typically $<110^\circ$.

Simulation Details. All GRIFFIN optimizations described here were carried out with NAMD 2.7¹⁵ with the CHARMM27 protein/lipid force field.²³ The lipid types used were palmitoyl oleyl phosphatidylcholine (POPC; 234 for c_{11} , 542 for K_{10} , and 532 for AcrB) and palmitoyl oleyl phosphatidylethanolamine (POPE; 470 for CaiT). The simulations were carried out at

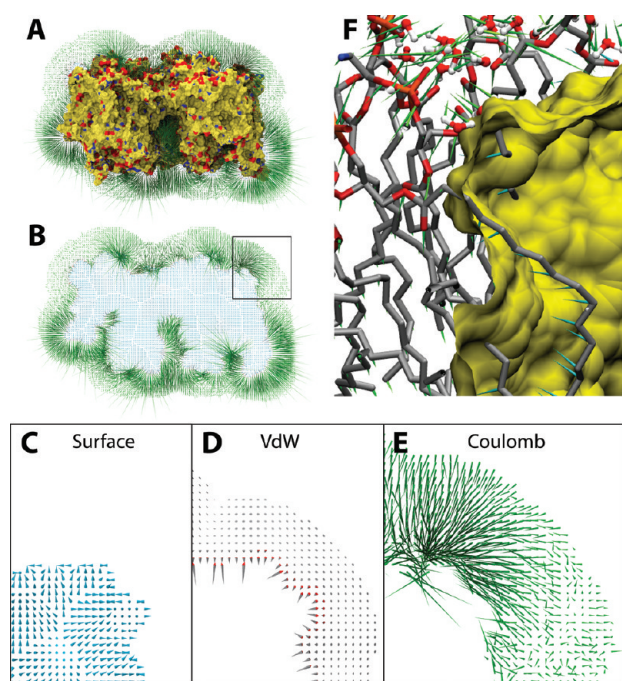


Figure 3. GRIFFIN force field for the CaiT protein. (A–E) The surface of CaiT (C atoms in yellow, O atoms in red, N atoms in blue) is shown overlaid on the corresponding GRIFFIN force field. Lines from a given grid point indicate the direction and magnitude of the forces generated by the protein atoms. These include surface forces (blue) inside the protein volume and electrostatic (green) and van der Waals attractive (gray) and repulsive (red) forces outside. (F) Actual forces derived by GRIFFIN inside (blue) and outside (green) the protein surface (yellow), on explicit lipid and water atoms, after interpolation and scaling of the force grid in (A)–(E).

constant temperature (298 and 310 K for POPC and POPE, respectively), using a Langevin thermostat (collision frequency of 100 ps^{-1}). The pressure was maintained along the membrane normal using a Nosé–Hoover Langevin piston barostat (1 atm), while the surface area of the membrane was kept constant. Electrostatic interactions were calculated using the particle mesh Ewald method (PME) with a real-space cutoff of 12 \AA . A cutoff distance of 12 \AA was also used for the van der Waals interactions. The MD integration time step was 1 fs; bonds involving hydrogen atoms were constrained with SETTLE. The number of lipid/water atoms in each optimization was 80 172 for c_{11} ring, 187 585 for K_{10} ring, 163 666 for CaiT, and 319 790 for AcrB.

RESULTS

GRIFFIN Forces Expel Lipid and Water Molecules from the Protein Volume. We began by testing the most basic GRIFFIN functionality, namely to gradually empty the volume of a membrane protein during the course of a MD simulation of a precarved, hydrated lipid bilayer. Our test system was the membrane rotor from a bacterial V-type ATPase (Figure 1A). This is a ring-shaped oligomeric assembly of 10 subunits, with a central pore that is plugged by lipid molecules.¹⁹ For this and subsequent tests, we employed NAMD as the MD engine coupled to GRIFFIN.

We monitor the progress of the procedure using two variables, namely the number of atoms inside the implicit protein volume

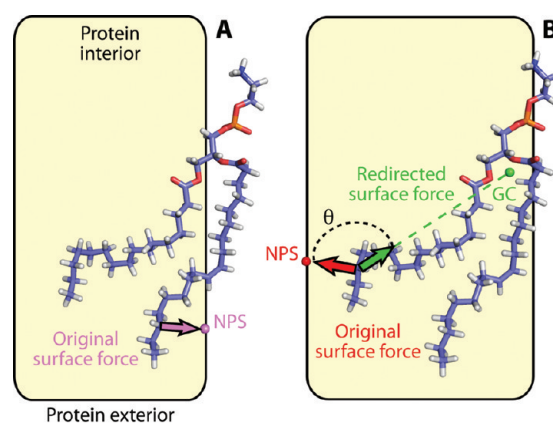


Figure 4. Schematic of force-redirection methodology. (A) Atoms in a lipid alkyl chain (sticks) entering the protein volume (yellow) experience surface forces toward the nearest point on the protein surface (NPS). (B) If the lipid chain spans most of the volume, however, atoms near the left-hand surface experience a surface force directed in the opposite direction from the rest of the chain (red). When the angle θ between the original surface force and the direction of the molecular geometric center (GC) exceeds a given threshold, the force is redirected toward the GC (green).

(excluding hydrogen atoms for convenience) and the maximum depth of any of those atoms beneath the protein surface. The K_{10} ring system starts out with about 6000 atoms inside the volume, which are up to about 10 \AA deep. As shown in Figure 5A, application of the surface forces calculated by GRIFFIN results in a rapid decline in the number of buried atoms, until a plateau is reached, reflecting a balance between the expelling forces and the external pressure from lipid and water. As expected, the rate of decrease and the plateau level are dependent on the strength of the applied surface forces; for example, with a force of $3.0 \text{ kcal/mol/\AA}^2$, the plateau is reached after $\sim 25 \text{ ps}$, having expelled ~ 3000 atoms. Subsequent stages in which the magnitude of the surface forces is increased (up to 9 kcal/mol/\AA^2 in this case) progressively empty the protein volume further. At the end of the last stage of our test, about 1300 atoms remain inside the implicit volume; however, these are within $\sim 1.5 \text{ \AA}$ of the protein surface (Figure 5B). As mentioned previously, this surface envelops the actual molecular surface with an offset equal to the probe radius (1.4 \AA); hence, at this point in the simulation the volume to be occupied by the protein is in fact empty of lipid or water atoms.

Avoiding Bidirectional Pulling and Trapping of Lipid Tails. The complexity of many membrane protein structures often implies that the curvature of their surface changes drastically in length scales comparable to a lipid molecule. Other features such as pores and gullies create multiple, noncontiguous protein–lipid interfaces. In these cases, a potential shortcoming of the surface-guided repulsion approach is that lipids may be trapped inside the protein volume. This occurs when two sets of atoms in a lipid molecule are pushed in opposite directions, each toward the nearest region of the protein surface.

To overcome this difficulty in GRIFFIN, we implemented a feature we refer to as “force redirection”. This means, roughly speaking, that atomic forces whose direction is opposite to the forces applied to the rest of the lipid molecule are redirected appropriately (see Methodology for a more precise description). To illustrate this feature, we employ a second rotor ring, this time a c_{11} oligomer from a bacterial F-type ATP synthase²⁰ (Figure 1B).

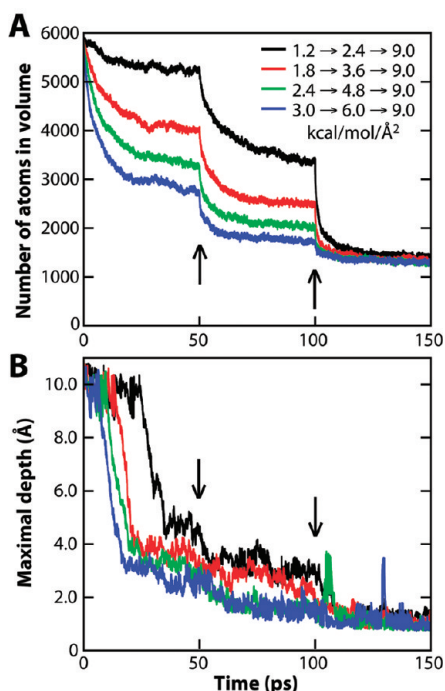


Figure 5. Progressive expulsion of phospholipid and water molecules from the volume of the K_{10} rotor ring due to GRIFFIN surface forces. (A) Number of atoms inside the protein volume, as a function of simulation time. The magnitude of the expelling surface forces calculated with GRIFFIN is increased in stages, at the time points marked with arrows. (B) Maximum depth inside the protein surface of any atom, as a function of simulation time. Hydrogen atoms are not considered in these or subsequent plots.

As shown in Figure 6A, several phospholipids in a membrane precarved for this protein span most of the distance between the inner and outer surfaces of the ring. Application of surface forces without the redirection functionality fails to expel these lipid tails from the protein volume; instead, as Figure 6B shows, the lipid tails are stretched in opposite directions. In this simple case, the situation can be somewhat resolved by increasing the magnitude of the surface forces (Figure 6D), as the lipid tail ultimately commits to one of the two surfaces. However, in general this will not be the case; moreover, large surface forces (>3 kcal/mol/Å²) should be avoided in the first stages of the procedure, as they otherwise result in unrealistic lipid-tail conformations, which subsequent simulations will be unlikely to correct.

By contrast, the redirection of problematic atomic forces rapidly resolves any conflicting lipid configurations, and allows GRIFFIN to empty the protein volume gradually and controllably (Figure 6C–E).

GRIFFIN Generates Lipid–Protein Interfaces with Interaction Specificity. An important improvement of GRIFFIN compared to its predecessor¹² is that it attains specificity in the protein–lipid interface, in addition to shape complementarity. To achieve this, the implicit protein force field overlaid on the explicit membrane system includes both electrostatic and van der Waals interactions; these act on lipid and water atoms outside the protein volume, in contrast to the surface forces, which only apply inside that volume (see Methodology; Figures 2 and 3).

Figure 7 illustrates both the degree of shape complementarity and the electrostatic specificity that may be attained with

GRIFFIN, using again the c_{11} rotor ring as an example. The repulsive surface-directed forces alone cause the lipid molecules to adapt neatly to the shape of the protein volume.¹² However, the electrostatic interactions in particular help to guide lipid groups to appropriate regions of the protein–membrane interface (Figure 7A cf. 7B), and thus reduce its overall electrostatic energy (Figure 7C). As illustrated in Figure 7D, during this optimization specific interactions are formed, for example, between the (explicit) polar head groups and oppositely charged atoms in the (implicit) protein. Arguably, that such interactions are already present in the starting configuration of a simulation is advantageous, as it implies that the subsequent equilibration of the system will be more realistic and efficient.

Geometric Objects for Exclusion or Inclusion of Specific Molecule Types in Desired Locations. Transmembrane pores and access pathways are characteristic features of membrane proteins that function as channels or transporters; many others contain large crevices in the lipid interface (Figure 1). Such cases present an additional challenge to the setup of a membrane protein simulation, since lipid molecules must be excluded from, e.g., hydrated pores, while water should be excluded from, e.g., lipid-filled crevices. To overcome these difficulties, GRIFFIN supports the use of geometric objects that define additional, molecule-type specific exclusion volumes. These objects are added to the protein volume during the initial membrane carving and/or during the subsequent MD/GRIFFIN simulation (see Methodology).

To illustrate this functionality, we consider as an example the trimeric carnitine/ γ -butyrobetaine antiporter CaiT from *Proteus mirabilis*,¹⁸ each protomer of which contains an aqueous vestibule leading toward the center of the membrane (Figures 1C and 8). Without using any additional exclusion volumes during the initial membrane carving, several lipid molecules are found inside these access pathways, amounting to ~ 160 atoms, not counting hydrogens (Figure 8A). Throughout a subsequent MD/GRIFFIN simulation—also without the objects—the majority of these atoms remain stubbornly in place (Figure 8B). If during the MD simulation, however, a spherical exclusion object is overlaid on each protomer so that it encompasses the access pathway (spheres, Figure 8), these regions are progressively emptied of lipid molecules (Figure 8, orange line). This result shows that the geometric objects are indeed integrated correctly into the protein volume, and that the expelling forces are calculated accordingly. Nevertheless, it should be noted that to use these additional exclusion volumes during the MD stage, but not the carving stage, implies that the overall area per lipid will be too small, particularly in the vicinity of the protein interface. Thus, it is recommended that the same objects be used also during the initial carving of the membrane. In our example, this excludes all but ~ 15 of the lipid atoms originally within those volumes (Figure 8C). At the end of the subsequent MD/GRIFFIN simulation, the access pathways in CaiT are entirely clear of lipid molecules (Figure 8D).

Performance and Portability. GRIFFIN is designed to be a stand-alone tool; i.e., it is intended to be compatible with any MD package, provided a minimal input/output interface. This interface, already included in, e.g., NAMD and GROMACS, allows for an external, accessory program (in this case GRIFFIN) to be executed at every time step of the simulation, to which the current coordinates of the molecular system are made available. If the program returns a set of atomic forces (and related quantities, e.g., energy, virial, etc), these will be added to the main algorithm and reflected in the simulation.

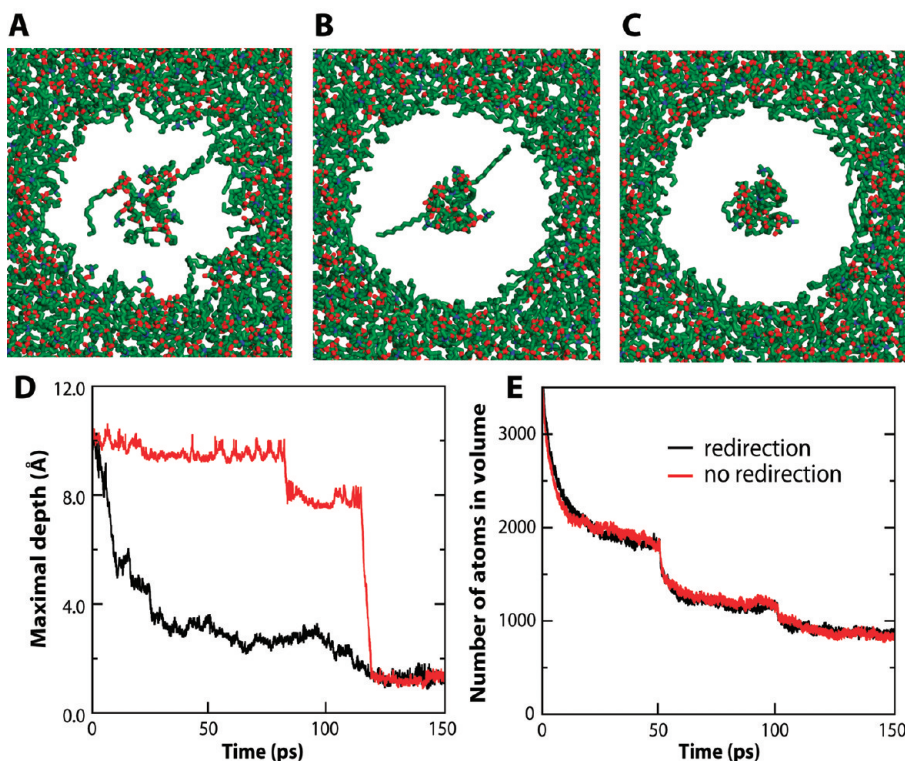


Figure 6. Effect of force redirection on the expulsion of lipid tails from the volume of the c_{11} rotor ring. (A–C) Several configurations of the lipids (green sticks) in the lipid bilayer, viewed along the normal to the membrane. Hydrogen and water atoms are omitted for clarity. Snapshots are taken (A) after carving, at $t = 0$ ps, and after the first stage of the GRIFFIN simulations, at $t = 100$ ps, either (B) without or (C) with force redirection. (D) The maximal depth inside the protein surface of any atom, as a function of simulation time. (E) Number of atoms inside the protein volume, as a function of simulation time. The magnitude of the forces was increased from 3.0 to 6.0 kcal/mol/Å² (at $t = 50$ ps) and 9.0 kcal/mol/Å² (at $t = 100$ ps). The force redirection angle was set to 110° (see Methodology).

As described previously, GRIFFIN precomputes a three-dimensional force grid before the start of the optimization phase, based on the membrane protein structure and its interactions with a probe particle positioned throughout the grid. During the GRIFFIN/MD simulation, the grid-point forces are interpolated and scaled to yield actual atomic forces. As mentioned, the initial force-grid calculation can be distributed among an adjustable number of computers, by portioning the grid accordingly. The computation of GRIFFIN atomic forces during run time can also be parallelized, using an MPI-compatible version of the program. As shown in Figure 9, the scalability of the atomic-force calculations is very good, especially for systems of about $\sim 100\,000$ atoms, such as the c_{11} ring. This scalability is, however, limited; the limit is imposed by the time required to exchange coordinates and forces with the MD software, which is a fixed overhead that depends on the computing infrastructure. Nevertheless, the total execution time, including the input/output of information, is still significantly reduced as the number of processor increases, at least within the range typically required. This allows GRIFFIN optimizations to be feasible within an affordable time frame even for very large systems such as AcrB, which comprises $\sim 300\,000$ atoms (Figure 10).

DISCUSSION

As new technologies lead to ever-growing computational power, and as algorithms improve in efficiency and accuracy, molecular dynamics simulations will be increasingly capable of

providing unique insights into the molecular mechanisms of membrane proteins. The strength of this theoretical tool—founded on statistical thermodynamics—lies in the atomic resolution at which the energetics and dynamics of the molecular system is simulated. These qualities, however, also pose a challenge in practice, as in most cases the structure of the environment of the membrane protein of interest is not known in atomic (or even molecular) detail. If this environment is modeled unrealistically to begin with, the results of the subsequent calculations will be questionable, as simulations are not designed to overcome large systematic errors. Therefore, it is generally advisable that a simulation research project comprises an initial stage focused not on the membrane protein but on the optimization of its lipid and solvent environment.

Current methodologies for preparing membrane protein simulations fall in two classes: those in which a lipid membrane is constructed around the protein of interest *de novo*, and those in which an existing lipid membrane model is somehow adapted. In the former class, one method uses a library of lipid conformations (derived from simulations of hydrated lipid bilayers) and progressively assembles individual molecules at designated locations around the protein; this is followed by a series of rigid-body rotations, translations, and energy minimizations.^{10,24–27} A second methodology in this class involves coarse-grained self-assembly simulations of a protein–lipid–water system, to which atomic detail is ultimately added.⁹

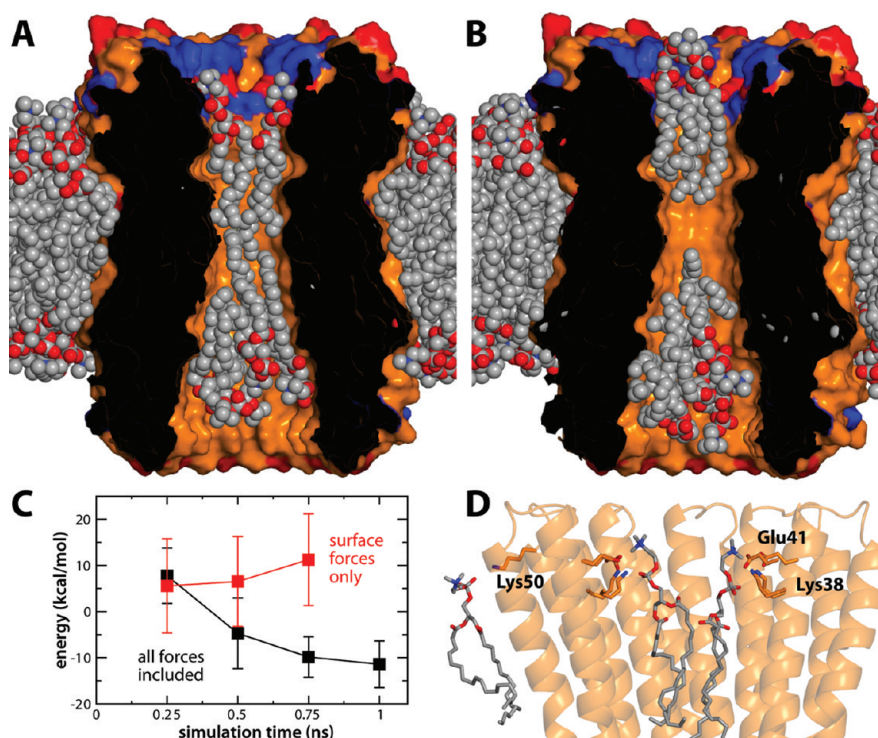


Figure 7. Shape and electrostatic complementarity between a hydrated lipid bilayer and the surface of the c_{11} rotor ring. (A) GRIFFIN-optimized interfaces between lipid (spheres) and protein (orange surface; basic and acidic residues in blue and red, respectively), after an extended GRIFFIN setup of the c_{11} ring. The system is viewed along the plane of the membrane at a cross section-through the center of the protein. Water molecules are omitted for clarity. (B) Same view as in (A), for an analogous GRIFFIN simulation in which only volume-exclusion forces were included. Note how the absence of electrostatic and van der Waals interactions with the protein causes the lipids in the inner pore to separate, driven by the gain in hydration by the surrounding waters. (C) Coulombic energy associated with interactions between charged side chains and lipid head groups, during the simulations in (A) and (B), shown as a function of simulation time. Note that no electrostatic complementarity develops when only volume-excluding forces are used. (D) Specific interactions between lipid head groups and charged side chains on the protein surface, in both the interior and the exterior of the ring, formed during the simulation in (A). The interaction distances, e.g., between amino and phosphate groups, approximately correspond to a hydrogen bond (<3.5 Å).

In the class of adaptive methodologies, the general aim is to embed the membrane protein structure within an existing hydrated lipid membrane. Two recent approaches employ coordinate scaling schemes to accomplish this. In the first of those, the membrane is expanded by scaling up the intermolecular lipid distances; the protein is then inserted into the space thus created and the membrane is recompressed gradually.¹¹ In the second, it is the protein structure that is compressed (onto a one-dimensional object perpendicular to the membrane plane), inserted into the lipid bilayer, and progressively expanded.¹³ A third methodology in this class, used widely, has been to employ the surface of the protein as a template with which to carve and shape a cavity within the hydrated membrane system, into which the protein itself may be then inserted.¹²

All these methodologies are in principle valid, but they have different strengths and weaknesses. An advantage of the adaptive approaches relative to the de novo methods is that they build upon the work carried out to optimize atomic-resolution molecular simulation models of lipid systems, be they single components or mixtures, and bilayers, bicelles, or nanodiscs. By contrast, de novo approaches require reconstruction of all of the interactions in the system, that is, not only those between the protein and its environment but also lipid–lipid and lipid–water interactions. This may be an important shortcoming given the extremely long time scales characteristic of lipid motions.

For example, lipid rotation about its long axis has a time scale of ~ 100 ns.²⁸ Additionally, the library-based approach may be too time-consuming for large membrane proteins, unless coarse-grained simulations of self-assembly are used, although these can be unpredictable and are somewhat arbitrary in restoring the atomic detail. Within the adaptive class, on the other hand, the scaling methods are more limited than de novo approaches in the case of complex membrane topologies, e.g., those with lipid-filled pores or gullies. Lastly, the surface-based approach is also suitable for any protein topology, while sharing the advantages of the other adaptive methods. Nevertheless, in the original version of this protocol the protein was represented by its shape alone, and thus only nonspecific interactions between protein and its environment could be optimized.¹² Another practical disadvantage of the original surface-based methodology is that it required software from several different sources (some unsupported) and also lacked automation.

The methodology presented here, referred to as GRIFFIN, builds on the surface-based adaptive approach, removing its original shortcomings and adding new features to expand its versatility and transferability. First, the preparation of the optimization stage (surface calculation, precarving of the hydrated membrane, etc.) has been automated and integrated into a single program. Also, it is now possible to refine the exclusion volume defined by the protein surface by adding geometric objects,

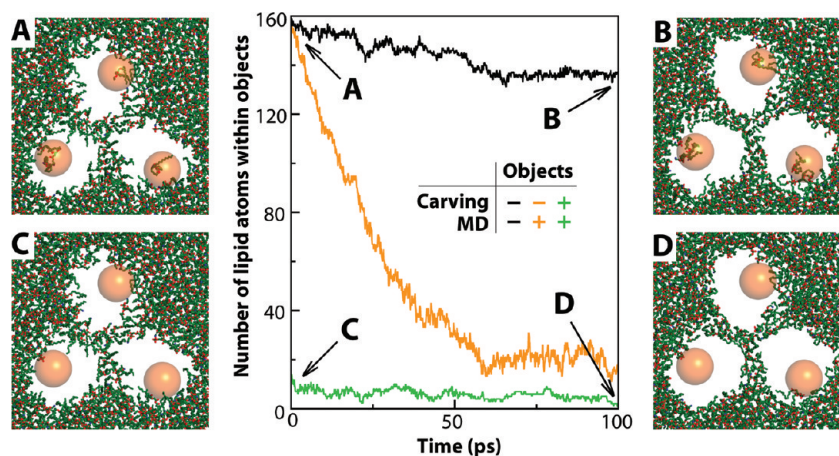


Figure 8. Use of geometric lipid-exclusion volumes to clear the aqueous substrate-access pathways in CaiT during GRIFFIN carving and MD simulations. The number of lipid atoms within three 22-Å-diameter spherical objects (spheres), positioned to encompass the access pathway in each of the three protomers, is plotted as a function of simulation time. Three cases are considered depending on whether the spherical exclusion volumes are used during the carving and the MD simulation stages. (A–D) The configuration of the lipid bilayer is depicted for (A) the precarved bilayer without objects; (B) at the end of the subsequent GRIFFIN/MD simulation, also without objects; (C) for the bilayer carved with objects; and (D) after subsequent GRIFFIN/MD simulations also using objects.

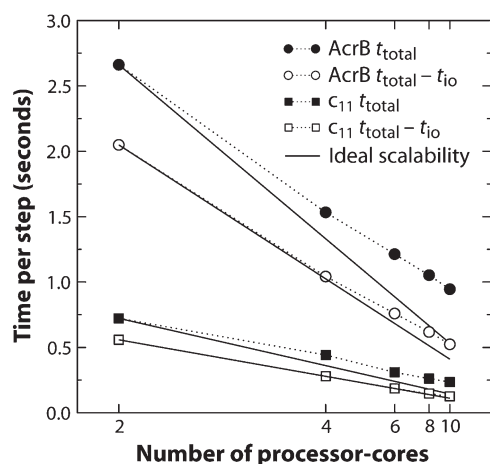


Figure 9. Performance and scalability of GRIFFIN, as a function of the number of processor-cores used. Calculation times per step are plotted for two membrane protein systems, namely the c_{11} ring and AcrB; both are embedded in a POPC membrane, comprising $\sim 80\,000$ and $\sim 320\,000$ atoms, respectively. Actual timings are shown with (t_{total}) and without ($t_{\text{total}} - t_{\text{io}}$) the overhead imposed by the input/output (i.e., coordinate/forces) exchange with the main MD program (dashed lines, symbols). These are compared with timings assuming ideal scalability relative to a two-core calculation (solid lines).

designed to be used as molecule type specific exclusion regions, e.g., for water-accessible pathways (as in CaiT; Figure 8) or lipid-filled cavities (as in AcrB or the c rings). In the actual optimization phase, which continues to involve a series of MD simulations, GRIFFIN now considers electrostatic and van der Waals interactions between protein and the hydrated membrane, alongside the expelling surface forces (Figures 2 and 3). These result in a much improved complementarity between the protein and its lipid and water environment, especially when the optimization is carried out gradually (Figure 7). This implies that subsequent simulations of the system will be more realistic and efficient. A lattice-based approach replaces the

original particle-based algorithm to meet the additional computational requirements due to the calculation of physical interactions during run time. To this end, a three-dimensional force grid including physical and exclusion forces is precomputed once, based on the protein structure (Figure 2). This implicit force field is overlaid on the explicit hydrated lipid membrane during the MD-based optimization; GRIFFIN thus derives actual atomic forces from the force grid at every step of the MD simulations and makes them available to the underlying MD engine (Figure 3F). Moreover, the code for the calculation of both the initial force grid and the time-dependent atomic forces is now written in C++ and parallelized; the scalability of the GRIFFIN force calculations is very good within the range of processor counts required in typical applications (Figure 9).

A practical caveat of the original surface-based method¹² was its limited compatibility, as it was integrated with a version of the MD software GROMACS that eventually became outdated. Therefore, the methodology has not been available to users of the recent versions of GROMACS or other MD packages. (Incidentally, lack of transferability among MD programs is a common limitation of the existing membrane-embedding methods mentioned above.) GRIFFIN, by contrast, is designed to be a stand-alone tool; it will work with any MD software that provides a minimal input/output (I/O) interface. This interface must be able to write out the coordinates of the explicit membrane system at every step; to execute a user-defined system command, e.g., to run GRIFFIN; and finally to read in and append a set of atomic forces (and related physical quantities) to the underlying MD algorithm. This procedure is clearly not optimal in terms of computational performance; however, as we have shown, the actual I/O overhead is not critical in practice even for systems of $\geq 100\,000$ atoms (Figure 9). Nevertheless, ongoing improvements are focused on the I/O protocol (and also on, e.g., memory management) since transferability is, in our view, paramount. At the present time, GRIFFIN can be readily employed alongside both NAMD 2.7¹⁵ and GROMACS v4.¹⁶ GRIFFIN and related materials will be made publicly available

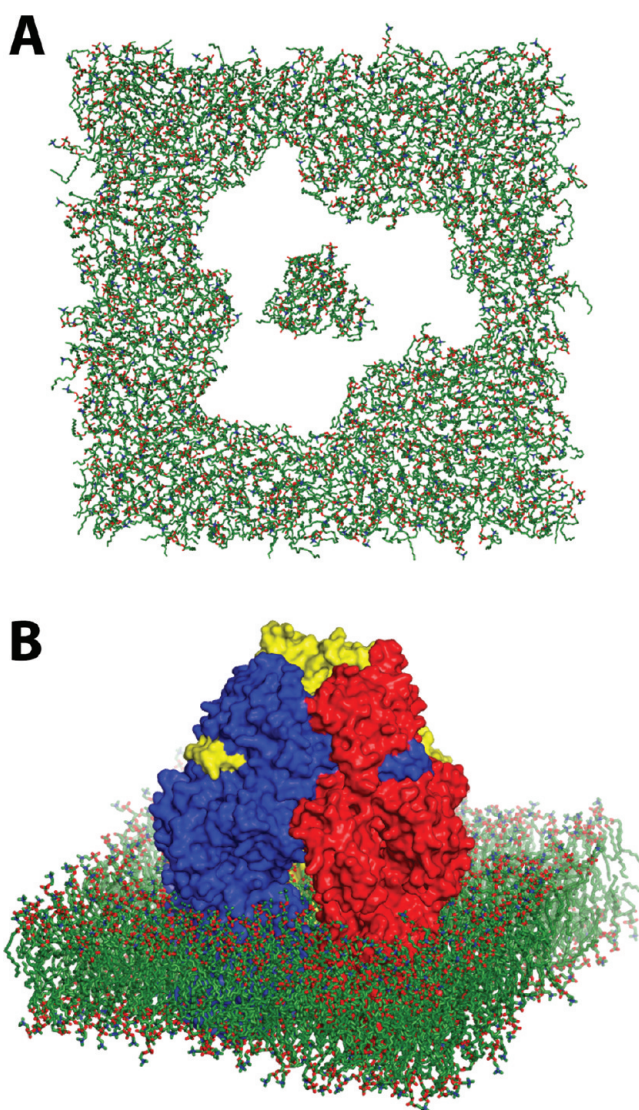


Figure 10. Molecular simulation model of AcrB in a lipid membrane. (A) The lipid membrane, following the GRIFFIN/MD optimization phase. (B) The complete system, after embedding and energy minimization of the actual AcrB structure into the GRIFFIN-optimized exclusion volume. Water is omitted for clarity.

for academic users at www.faraldolab.org and www.forrestlab.org.

CONCLUSIONS

We have introduced a versatile and efficient methodology to prepare molecular dynamics simulations of membrane proteins, specifically aimed at optimizing the interface between the protein structure and its lipid environment. This methodology, based on a new grid-based simulation tool named GRIFFIN, is uniquely suited for membrane proteins of intricate topologies and irregular interfaces, such as those increasingly found among channels and transporters. Another advantage of GRIFFIN is that it builds upon already optimized simulation models of lipid–solvent systems, be they simple homogeneous bilayers, multicomponent membranes, or even nanodiscs. Lastly, GRIFFIN is a stand-alone tool that is designed to work alongside any existing molecular dynamics package, such as NAMD or GROMACS.

AUTHOR INFORMATION

Corresponding Author

*Phone: +49 69 6303 1600 (L.R.F.), +49 69 6303 1500 (J.D.F.-G).
Fax: +49 69 6303 1502 (L.R.F.), +49 69 6303 1502 (J.D.F.-G).
E-mail: lucy.forrest@biophys.mpg.de (L.R.F.), jose.faraldo@biophys.mpg.de (J.D.F.-G).

ACKNOWLEDGMENT

We thank Gerrit Groenhof (MPI of Biophysical Chemistry), for his assistance with the GROMACS interface for GRIFFIN, and Wenchang Zhou (University of Konstanz), for his involvement in the AcrB simulations. This work was supported in part by the DFG Collaborative Research Center 807 “Transport and Communication across Biological Membranes” (R.S. and L.R.F.), the DFG Cluster of Excellence “Macromolecular Complexes” (J.D.F.-G.), and the Behrens-Weise-Stiftung (C.A.). Computational resources were in part provided by the Jülich Supercomputing Center.

REFERENCES

- (1) Wallin, E.; von Heijne, G. Genome-wide analysis of integral membrane proteins from eubacterial, archean, and eukaryotic organisms. *Protein Sci.* **1998**, *7*, 1029–1038.
- (2) Krogh, A.; Larsson, B.; von Heijne, G.; Sonnhammer, E. L. L. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J. Mol. Biol.* **2001**, *305*, 567–580.
- (3) Gerstein, M. A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *J. Mol. Biol.* **1997**, *274*, 562–576.
- (4) White, S. H. Biophysical dissection of membrane proteins. *Nature* **2009**, *459*, 344–346.
- (5) Ozcan, N.; Ejsing, C. S.; Shevchenko, A.; Lipski, A.; Morbach, S.; Kramer, R. Osmolality, temperature, and membrane lipid composition modulate the activity of betaine transporter BetP in *Corynebacterium glutamicum*. *J. Bacteriol.* **2007**, *189*, 7485–7496.
- (6) Swartz, K. J. Sensing voltage across lipid membranes. *Nature* **2008**, *456*, 891–897.
- (7) Vasquez, V. A structural mechanism for MscS gating in lipid bilayers. *Science* **2008**, *321*, 1210–1213.
- (8) Watt, I. N.; Montgomery, M. G.; Runswick, M. J.; Leslie, A. G. W.; Walker, J. E. Bioenergetic cost of making an adenosine triphosphate molecule in animal mitochondria. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 16823–16827.
- (9) Scott, K. A.; Bond, P. J.; Ivetac, A.; Chetwynd, A. P.; Khalid, S.; Sansom, M. S. P. Coarse-grained MD simulations of membrane protein–bilayer self-assembly. *Structure* **2008**, *16*, 621–630.
- (10) Woolf, T. B.; Roux, B. Molecular dynamics simulation of the gramicidin channel in a phospholipid bilayer. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 11631–11635.
- (11) Kandt, C.; Ash, W. L.; Peter Tieleman, D. Setting up and running molecular dynamics simulations of membrane proteins. *Methods* **2007**, *41*, 475–488.
- (12) Faraldo-Gómez, J. D.; Smith, G. R.; Sansom, M. S. P. Setting up and optimization of membrane protein simulations. *Eur. Biophys. J.* **2002**, *31*, 217–227.
- (13) Wolf, M. G.; Hoefling, M.; Aponte-Santamaría, C.; Grubmüller, H.; Groenhof, G. *g_membed*: Efficient insertion of a membrane protein into an equilibrated lipid bilayer with minimal perturbation. *J. Comput. Chem.* **2010**, *31*, 2169–2174.
- (14) Shen, L.; Bassolino, D.; Stouch, T. Transmembrane helix structure, dynamics, and interactions: multi-nanosecond molecular dynamics simulations. *Biophys. J.* **1997**, *73*, 3–20.

(15) Phillips, J.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R.; Kale, L.; Schulten, K. Scalable molecular dynamics in NAMD. *J. Comput. Chem.* **2005**, *26*, 1781–1802.

(16) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GRO-MACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.

(17) 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. CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **2009**, *30*, 1545–1614.

(18) Schulze, S.; Koester, S.; Geldmacher, U.; Terwisscha van Scheltinga, A. C.; Kuehlbrandt, W. Structural basis of cooperative substrate binding and Na⁺-independent transport in the carnitine/butyrobetaine antiporter CaiT. *Nature* **2010**, *467*, 233–237.

(19) Murata, T.; Yamato, I.; Kakinuma, Y.; Leslie, A. G. W.; Walker, J. E. Structure of the Rotor of the V-Type Na⁺-ATPase from *Enterococcus hirae*. *Science* **2005**, *308*, 654–659.

(20) Meier, T.; Krahe, A.; Bond, P. J.; Pogoryelov, D.; Diederichs, K.; Faraldo-Gómez, J. D. Complete ion-coordination structure in the rotor ring of Na⁺-dependent F-ATP synthases. *J. Mol. Biol.* **2009**, *391*, 498–507.

(21) Seeger, M. A.; Schiefner, A.; Eicher, T.; Verrey, F.; Diederichs, K.; Pos, K. M. Structural asymmetry of AcrB trimer suggests a peristaltic pump mechanism. *Science* **2006**, *313*, 1295–1298.

(22) Connolly, M. L. Analytical molecular surface calculations. *J. Appl. Crystallogr.* **1983**, *16*, 548–558.

(23) MacKerrell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCartney, 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.; Wiorcikiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(24) Woolf, T.; Roux, B. Structure, energetics, and dynamics of lipid-protein interactions: A molecular dynamics study of the gramicidin A channel in a DMPC bilayer. *Proteins: Struct., Funct., Genet.* **1996**, *24*, 92–114.

(25) Roux, B.; Woolf, T., Molecular dynamics of Pfl coat protein in a phospholipid bilayer. In *Biological Membranes*; Merz, K. M. J., Roux, B., Eds.; Birkhauser: Boston, 1996; pp 555–587.

(26) Petrache, H.; Grossfield, A.; MacKenzie, K. R.; Engelman, D.; Woolf, T. B. Modulation of glycophorin A transmembrane helix interactions by lipid bilayers: molecular dynamics calculations. *J. Mol. Biol.* **2000**, *302*, 727–746.

(27) Jo, S.; Kim, T.; Im, W. Automated builder and database of protein/membrane complexes for molecular dynamics simulations. *PLoS One* **2007**, *2*, e880.

(28) Klauda, J. B.; Roberts, M. F.; Redfield, A. G.; Brooks, B. R.; Pastor, R. W. Rotation of lipids in membranes: molecular dynamics simulation, ³¹P spin-lattice relaxation, and rigid-body dynamics. *Biophys. J.* **2008**, *94*, 3074–3083.

The Catalytic Mechanism of RNA Polymerase II

Alexandra T. P. Carvalho, Pedro A. Fernandes, and Maria J. Ramos*

Requimte, Faculty of Sciences of Porto, Rua do Campo Alegre S/N, 4169-007 Porto, Portugal

S Supporting Information

ABSTRACT: Eukaryotic RNA polymerase II (RNAP II) transcribes the DNA into mRNA. The presence of two metal ions (usually Mg^{2+}) and conserved aspartate residues in the active sites of all nucleic acid polymerases led to the adoption of a universal catalytic mechanism, known as the “two metal ion catalysis”. In this scheme, it is assumed that the coordination shell of Mg^{2+} (geometry, number, and identity of the ligands) is basically the same for all of the enzymes, despite the significant differences in sequence and structure commonly found in multisubunit RNA polymerases versus single-subunit RNA polymerases and DNA polymerases. Here, we have studied the catalytic mechanism of RNAP II and found very interesting variations to the postulated mechanism. We have used an array of techniques that included thermodynamic integration free energy calculations and electronic structure calculations with pure DFT as well as hybrid DFT/semiempirical methods to understand this important mechanism. We have studied four different catalytic pathways in total, resulting from different combinations of proton donors/acceptors for the two proton transfers experimentally detected (deprotonation of the 3' hydroxyl of the terminal nucleotide (HO_{RNA}) and protonation of pyrophosphate). The obtained data unambiguously show that the catalytic mechanism involves the deprotonation of HO_{RNA} by a hydroxide ion coming from the bulk solvent, the protonation of pyrophosphate by the active site His1085, and the nucleophilic attack to the substrate by O^-_{RNA} . The overall barrier is 9.9 kcal/mol. This mechanism differs from those proposed in the identity of the general acid. The deprotonation of the HO_{RNA} and the transition state for the nucleophilic attack are similar to some (but not all) of the family members.

1. INTRODUCTION

Two different evolutive solutions have emerged for transcription: The first corresponds to a large family of multisubunit RNAPs, which includes bacterial enzymes and eukaryotic nuclear enzymes (RNAP I, RNAP II, and RNAP III), among others. The second involves single-subunit RNAPs that include enzymes from bacteriophages, such as T7 RNAP, among others. Single-subunit RNAPs share a strong structural homology to DNA polymerases (DNAPs).

This work deals with RNAP II, which is the multisubunit eukaryotic RNA polymerase responsible for the transcription of the genomic DNA into mRNA in eukaryotic cells. It consists of a 10-subunit catalytic core and a heterodimeric Rbp4/7 subcomplex. The active site is located in the interface between the core subunits Rpb1 and Rpb2 and contains two Mg^{2+} ions (Figure 1). According to Wang et al. for RNAP II, one is a persistently bound Mg^{2+} (the catalytic magnesium, Mg_A^{2+}).¹ The second Mg^{2+} (the nucleotide-binding magnesium, Mg_B^{2+}) is only found in the transcribing complex with the substrate. There is not a general mechanism for how Mg^{2+} binding happens during the catalytic cycle of polymerases, and such a general mechanism may not exist at all. In fact, for DNA polymerase β , both ions are believed to leave the active site after the catalytic reaction and to re-enter for the new cycle.

The ions are held in place through coordination to four aspartates (Asp481, Asp483, Asp485, and Asp837) and the triphosphate of the substrate.¹ The enzyme has a very flexible motif at the active site (the trigger loop) that closes the active site upon substrate (NTPs) binding.¹

The four-step cyclic process of nucleotide addition first involves the selection of a specific NTP. This occurs in two

phases, with an initial binding to the entry site, in an inverted orientation, followed by rotation to the nucleotide addition site, for pairing with the complementary template DNA. In the second step, the very flexible trigger loop suffers a conformation change and closes the active site. The catalyzed phosphodiester bond formation follows, and finally, translocation finishes the cycle.

The reaction mechanism is postulated to be common to all nucleic acid polymerases. It involves a nucleophilic attack by O^-_{RNA} to the phosphorus atom of the α phosphate of the substrate (P_α), forming a new phosphodiester bond and releasing pyrophosphate (PPi; see Schemes 1 and 2). For this reaction to be accomplished, two proton transfers must also take place: HO_{RNA} must be deprotonated, and PPi must be protonated, as it is known to be released in the monoprotonated form.^{2,3} The general scheme does not explain the first proton transfer. The protonation of PPi is also not addressed in the general two ion catalytic scheme. Therefore, the HO_{RNA} proton acceptor and PPi proton donor are presently unknown. Three hypotheses are advanced for the identity of the base that deprotonates HO_{RNA} in DNA polymerases (DNAPs). The first corresponds to a suitable protein residue, namely, the catalytic triad Asp closer to the DNA terminus.^{4–6} In RNAP II, it corresponds to Asp485. In the second hypothesis, the base is one of the two nonbridging α -oxygens of the substrate ($O\alpha$).^{7,8} The third hypothesis is a HO^- ion coming from the bulk solution (the external HO^- hypothesis).^{6,9,10} This last hypothesis was first proposed for the

Received: October 8, 2010

Published: March 18, 2011

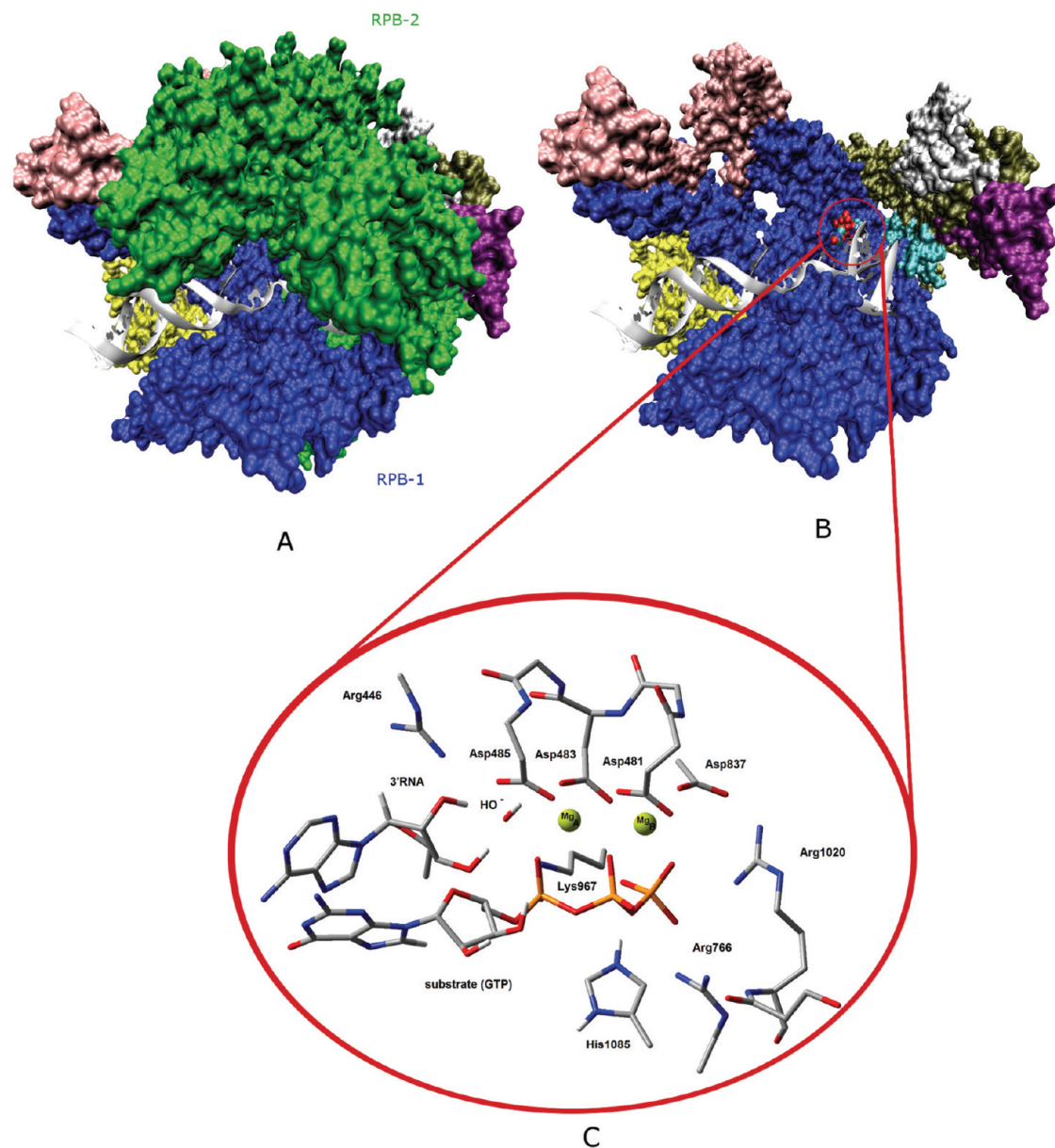
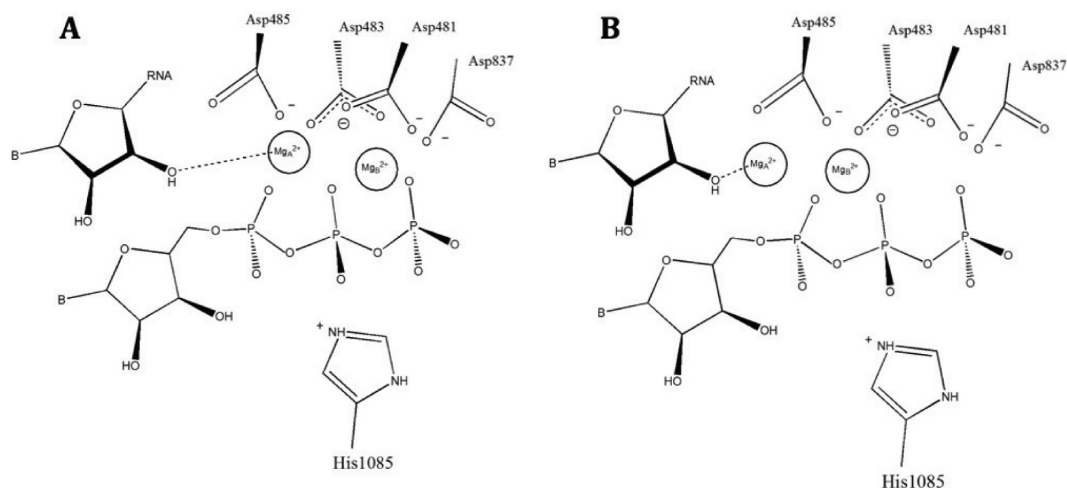


Figure 1. The RNAP II structure. (A) The overall complex protein. DNA:RNA:substrate. (B) The same structure with the Rpb-2 deleted so that the active site (VDW representation) and DNA:RNA hybrid can be seen. (C) The atoms of the active site included in the DFT calculations. Each model contains the substrate (GTP); the 3' terminal nucleotide of the RNA chain; the two magnesium ions; the aspartates 481, 483, 485, and 837; the protonated histidine 1085; the arginines 446, 766, and 1020; and Lys967. Some of the models also included a HO^- ion. Most of the hydrogen atoms were deleted in the figure for simplicity.

exonuclease reaction of DNA polymerase I¹¹ and later by Flórian et al.⁶ for the polymerization reaction by T7 DNA polymerase. Flórian et al. estimated that the free energy required for the hydroxide to deprotonate the HO_{RNA} at neutral pH was in the range of 1.5–3.0 kcal/mol.¹⁰ On the basis of the similarity between the active sites of RNAP II and DNAPs, it seems logical to consider the same hypothesis for the deprotonation of HO_{RNA} .

Recently, the protonation of the leaving PPi by a protein residue was proposed.² This assumption was based on a proton-inventory experiment. The data were consistent with a model in which more than one proton transfer occurred during a nucleotidyl transfer reaction. Subsequently, to check this hypothesis,

nucleotide incorporation rates were measured for four representatives of the four classes of nucleic polymerases (the RNA-dependent RNA polymerase from poliovirus (PV), the reverse transcriptase (RT) from human immunodeficiency virus type 1, the DNA-dependent DNA polymerase from bacteriophage RB69, and the DNA-dependent RNA polymerase from bacteriophage T7). It was concluded that the polymerases employ a general acid catalysis mechanism for nucleotidyl transfer. Importantly, the general acid is not absolutely essential but provides a 50–2000 fold rate enhancement depending upon the polymerase evaluated.³ Multisubunit RNA polymerases contain a histidine in the trigger loop that can serve as a general acid.¹ Changing His1085 to Tyr reduces the rate of catalysis by an order

Scheme 1. The Position of the Mg^{2+} Ions in the MD Simulations (A) and in the Two Metal Ions Catalysis Scheme (B)

of magnitude without changing the observed affinity for the nucleotide substrate,¹² which is fully consistent with the above-mentioned results and points to His1085 as a candidate for the role of acid in the general acid catalysis mechanism.

The two active site Mg^{2+} ions participate actively in catalysis. The two-metal ion catalysis scheme was first described 20 years ago, upon the determination of the crystallographic structure of the exonuclease active site of DNA polymerase I.¹³ The hypothesis was later applied to RNA splicing, hydrolysis of tRNA by RNase P, and several DNA polymerases.^{14,15} All of these mechanistic proposals were essentially based on the observation of the crystallographic structures of the mentioned enzymes. Mg_A^{2+} is thought to orient and activate the nucleophile (here, HO_{RNA}). Activation is achieved through the lowering of the $\text{p}K_a$ of the nucleophile, facilitating its deprotonation. Mg_B^{2+} is thought to stabilize the negative charge that develops in the postulated pentacoordinated transition state and facilitate the leaving of PPi .¹⁴

If these mechanisms are correct, this would be a clear case of convergent evolution. A myriad of structurally unrelated enzymes, involved in phosphoryl transfer reactions, would employ the same two-ion mechanism to accomplish their roles.

The ions seem to be essential for substrate recognition and for the greater specificity of the enzymes. The reported distance between the metal ions in most polymerases is around 4 Å. However the positions and distance of the metals may change at each reaction step due to differences in the coordination environment. Yang et al. hypothesized that the conformation with the metals separated by 4 Å should represent a “resting state” and that during the reaction Mg_A^{2+} should move toward Mg_B^{2+} , bringing the nucleophile within striking distance for phosphoryl bond formation.¹⁶ The approximation of the Mg^{2+} ions could also better neutralize the developing negative charge on the pentacoordinated intermediate.¹⁶

The purpose of the present study is to identify the elementary steps that constitute the catalytic pathway, quantifying the activation barriers and reaction energies, characterizing the transition states, and finally elucidating and understanding the mechanism by which RNAP II catalyzes the biosynthesis of the mRNA.

To fulfill our goal, we have used several techniques, namely molecular dynamics (MD), thermodynamic integration (TI) free

energy calculations, density functional theory (DFT), and hybrid quantum mechanics/molecular mechanics (QM/MM) methods, more precisely DFT:PM3MM.

We have explored four different mechanistic hypotheses that differed mostly in the proton transfer reactions. The results allowed for the identification and understanding of the catalytic mechanism of RNAP II, which is in fact different from those proposed in the literature as far as the identity of the general acid is concerned. Furthermore, the protonation of the HO_{RNA} and the transition state for the nucleophilic attack are similar to some, albeit not all, of the family members.

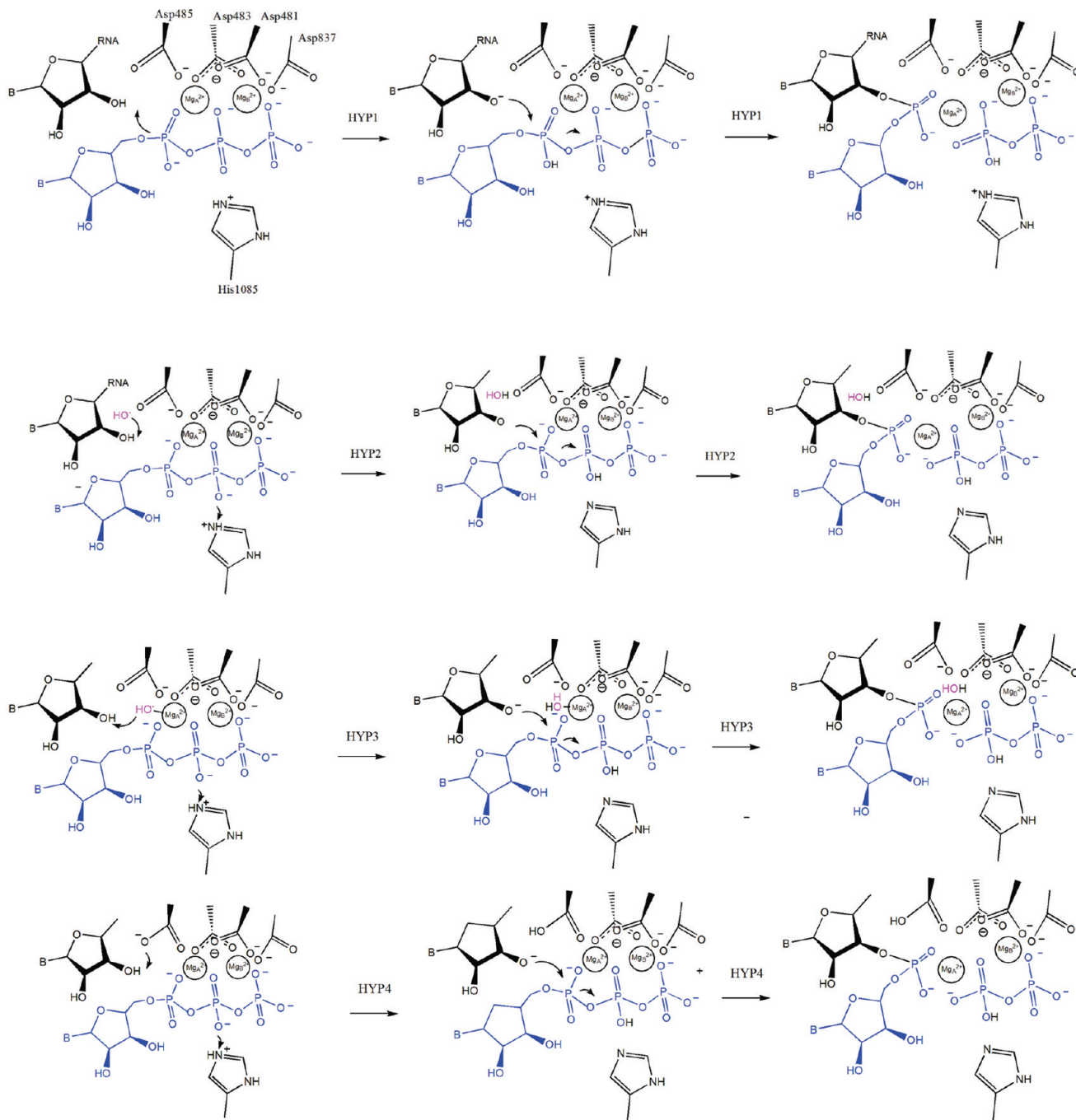
2. METHODOLOGY

2.1. Molecular Dynamics. The molecular dynamics simulations were performed with the Groningen Machine for Chemical Simulations, version 4, Gromacs 4, with the amber99 force field [11].

The models were built using the coordinates of the crystallographic structure of RNAP II from *S. cerevisiae* with PDB code 2E2H.¹ This structure is the most recent structure of RNAP II with the substrate at a low Mg^{2+} concentration. The resolution is 3.95 Å, which is far from excellent but is accurate enough for the study conducted here. The missing loops were modeled resorting to the structures 2NVZ and 2E2I.

The full enzyme was first neutralized by adding 83 Na^+ ions and solvated in a cubic box with 181 875 waters (which corresponds to a distance of at least 10 Å besides the protein atoms). The geometry was optimized with 400 steps of steepest descent minimization. All subsequent simulations were carried out in the NTP ensemble with periodic boundary conditions. The Berendsen temperature coupled with velocity rescaling was employed to maintain a constant temperature at 300 K. The pressure was kept constant at 1 atm with a Parrinello–Rahman barostat. Isotropic position scaling was used to maintain the pressure with a relaxation time of 0.5 ps. The particle mesh Ewald (PME) method was employed to compute the electrostatic interactions with a cutoff of 0.8 nm. LINCS constraints were applied to all bonds involving hydrogen. The time step was 0.002 ps. The trajectory was saved every 1 ps. We performed a total of 50 ps of equilibration and 20 ns of a production run.

Scheme 2. Intramolecular Proton Transfer Followed by Nucleophilic Attack (HYP1), the External HO⁻ Hypothesis (HYP2), the Transfer of the HO_{RNA} Proton to an HO⁻ Coordinated to Mg_A²⁺ (HYP3) and the Transfer of the HO_{RNA} Proton to the Conserved Protein Residue Asp485 (HYP4)



2.2. QM Models. The models included all species relevant for the mechanism, i.e., the substrate (GTP); the 3' terminal nucleotide of the bound RNA; the two Mg²⁺ ions; the catalytic triad (Rpb1-Asp481, 483, and 485); the side chain of Asp837 up to the β carbon (which is also in the coordination sphere of Mg_B²⁺); the side chain of His1085 from Rpb1 (which may act as a general acid in the mechanism); and the side chains of Arg446, Arg766, Arg1020, and Lys967 from Rpb2. These latter residues were included because they engage in H bonds with the oxygen atoms of the triphosphate (Figure 1). A HO⁻ ion was also

included when its participation in the catalytic cycle was hypothesized. The models included 226 atoms in total (228 with HO⁻). The total charge of the models was 0 in the cases where the HO⁻ was included and +1 in the remaining.

The calculations were done with Gaussian 03. We have used the ONIOM method to optimize the geometries. We have used a QM/QM methodology with the high level layer treated with DFT and the remaining with PM3MM. In these instances, the charge transfer is accounted for at the lower level of theory. The subtractive ONIOM method has the same limitations of any

other additive hybrid method within this application, which are related to the less accurate description of the lower level region and interactions between layers. These limitations were greatly alleviated in the final energy calculations, which were done with the full system described at a high level (DFT).

The higher level layer contains the triphosphate of the substrate (GTP), the ribose of the RNA, the two magnesium ions, the δ carbon and the two δ oxygens of the four aspartate residues, His1085 (this last in all cases except in the study of HYP1, as the His here does not participate in the reaction), and the HO^- ion.

A recent study¹⁷ has shown that MPWB1K was the most accurate functional for the description of the hydrolysis of phosphodiester bonds. In the same study, it was concluded that the geometries calculated with B3LYP have negligible differences in relation to the MPWB1K geometries. Concerning the energies, B3LYP was shown to overestimate the barriers by 2.94 kcal/mol and overestimate the reaction energies by 2.02 kcal/mol; given such small differences and the precise knowledge of the inaccuracies involved, we opted to use B3LYP, as this functional is much more numerically stable than the ones that explicitly depend on the kinetic energy density. Numeric stability is crucial in facilitating calculations in a system of this size.

The geometries were optimized at the ONIOM (B3LYP/6-31G(d):PM3MM) level. The energies were calculated with the whole model at the DFT level (B3LYP/6-311++G(2d,2p)). This combination of theoretical levels has been shown in the past to provide geometries and energies accurate enough for the purpose in mind.^{18,19} All stationary points were optimized without constraints.

Frequency calculations were carried out, with the resulting number of imaginary frequencies confirming the nature of the stationary points. The zero point energy and the thermal and entropy contributions for the Gibbs free energy were calculated at the same theoretical level. The electrostatic effect of the remaining protein was accounted for using a dielectric continuum (PCM) with a dielectric constant of 4.

Although the active site of the X-ray structure of RNAP II of Wang et al.¹ (with the substrate before the addition) is similar to some of the reported active sites of DNAPs,^{6,9,20} there are some important differences. The most striking is that Mg_A^{2+} is not positioned between HO_{RNA} and the substrate O_ω , but instead well above P_α (Figure 1).

2.3. Thermodynamic Integration. We have simulated the transfer of a HO^- ion from bulk solvent to the protein active site. This was done with the thermodynamic integration (TI) technique, through the annihilation of the ion in bulk solvent and creation of the ion in the proper active site place. The process was carried out in two separate stages, neutralization of the atomic charges and annihilation of the van der Waals parameters. Due to the large size of the enzyme, we have used a sphere with a radius of 25 Å, centered in the HO_{RNA} group.

The TI method allows for computation of the free energy difference between two states by gradually transforming the initial state into the final state. The parameter λ represents the state of the system along the transformation (eq 1):

$$\Delta G_{\text{TI}} = \int_0^1 \left\langle \frac{\partial H(\lambda)}{\partial \lambda} \right\rangle_\lambda d\lambda \quad (1)$$

$H(\lambda)$ is the Hamiltonian for state λ . λ equals zero for the initial state and 1 for the final state. The brackets represent an ensemble

average. As the exact calculation of ΔG_{TI} would require an infinite number of ensemble averages for λ , ranging continuously from 0 to 1, the following approximation is employed:

$$\Delta G_{\text{TI}} = \sum_i \left\langle \frac{\partial H(\lambda)}{\partial \lambda} \right\rangle_\lambda \Delta \lambda_i \quad (2)$$

The transformation simulated here (the annihilation of HO^- in aqueous solution and the creation of HO^- in the active site) was performed in two stages with nine λ points each. The first stage, $\lambda = 0.1$, corresponds to the HO^- ion in water (or in the active site), and $\lambda = 0.9$ corresponds to the ion at the same place but without atomic charges (neutralization). In the second stage, we started from the neutral ion ($\lambda = 0.1$) and gradually reduced its van der Waals interactions until $\lambda = 0.9$, where the ion was fully annihilated. At each λ point, we have performed 500 steps of steepest descent minimization, followed by 50 ps of equilibration and 200 ps of a production run. The whole calculation was 4.5-ns-long.

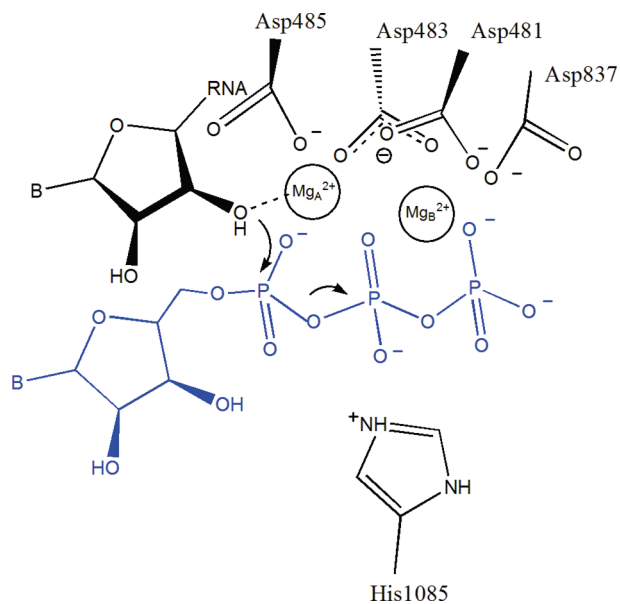
3. RESULTS AND DISCUSSION

Before discussing in detail the mechanistic pathways that we have explored, we will give an overview of the experimental and theoretical backgrounds that made us assume the specific coordination shells of Mg_A^{2+} used here. In the crystallographic structure of RNAP II, Mg_A^{2+} is pentacoordinated and Mg_B^{2+} is tetraordinated. However, these coordination shells must be interpreted with caution, as the terminal RNA nucleotide misses the HO_{RNA} group (to prevent the attack to the substrate). Therefore, the RNA is necessarily unbound from the Mg_A^{2+} . This is a general limitation present in all crystallographic files of RNAPs and DNAPs. All of those that contain RNA or DNA and the respective substrate must contain a chemical modification to prevent the reaction, usually the deletion of the 3' hydroxyls of the terminal RNA or DNA nucleotides. This modification influences the metal coordination shell. Therefore, it has not been possible to show if many of these enzymes (and specifically RNAP II) indeed have the HO_{RNA} bound to Mg_A^{2+} at the beginning of the cycle.

Several experimental and computational studies give support for a coordination number for Mg^{2+} in the range of four to six.²¹ In aqueous solutions, Mg^{2+} was almost always observed hexacoordinated. In 1999, gas phase electrospray ionization experiments revealed two different $[\text{Mg}(\text{H}_2\text{O})_6]^{2+}$ isomeric forms.²² DFT calculations have shown that in one of these forms Mg^{2+} had a pentacoordinated structure, with the sixth water molecule in the second coordination shell. Pentacoordinated catalytic Mg^{2+} ions were also observed in the X-ray structures of Mg^{2+} containing enzymes such as the Klenow fragment of DNA polymerase I,¹³ the ϵ subunit of *E. coli* DNA polymerase,²³ the SRP GTPase Ffh,²⁴ ribonuclease H,²⁵ and tryptophanyl-tRNA synthetase.²⁶

In a DFT study on the influence of the identity of the ligands on the coordination number of Mg^{2+} , Kluge and Weston found that the presence of one hydroxide ligand triggers a change from the typically observed octahedral geometry to a trigonal bipyramidal geometry (with five ligands).²⁷ When two hydroxide ions are present, the tetrahedral coordination geometry is preferred. However water, carboxylate, or ammonia ligands were not able to change the preference of Mg^{2+} for the octahedral geometry. Kluge and Weston pointed out that the reason for this preference is electrostatic. However they were unable to find out the actual reasons why HO^- stabilizes the pentacoordinated geometry.

Scheme 3. The CHECK Model



Consequently, other highly charged ligands such as the triphosphate of GTP may also induce a pentacoordinated geometry.

We have performed MD simulations on complexes of RNAP II with RNA/DNA and the substrate in explicit solvent. We have performed three MD runs, two of them without any constraint (besides the bond lengths involving hydrogen atoms). In both of these runs, the relative positions of the magnesium ions have not changed in relation to the X-ray structure: Mg_A^{2+} is maintained close to the oxygens of P_ω and Mg_B^{2+} is kept between P_β and P_γ oxygens.

In the third run, we constrained the position of specific atoms to obtain the configuration that is considered to be more typical for two metal ion catalysis in other enzymes: (i) The distance between the RNA terminus and the Mg_A^{2+} was constrained to 2 Å. (ii) The distance between the two metal ions was kept at 4 Å. (iii) We placed one Mg^{2+} on each side of oxygen 1 α (oxygen of the scissile phosphorus).

We ran the constrained simulation for 1 ns. Afterward, we released the constraints and ran the simulation for an additional 20 ns. When we released the constraints, the atoms changed almost immediately to positions very similar to the ones in the simulations without constraints.

One of the possible explanations for the particular cofactor arrangement in this enzyme is that the active site contains four negative Asp residues (not three Asp/Glu as in DNA polymerases). The position of the Mg^{2+} ions in the two-metal-ion catalysis scheme would create a charge imbalance at P_γ , due to the additional Asp present in RNAP (please see Scheme 1B). For that reason, it can be easily understood that the Mg^{2+} ions adopt the positions reported in Scheme 1A in all simulations.

After making conclusions about the positions of the Mg^{2+} ions, we then wanted to check the position of the HO_{RNA} in relation to Mg_A^{2+} since this was not clear from the MD simulations. In some simulations, the HO_{RNA} was strongly coordinated to Mg_A^{2+} and in others was weakly coordinated. We started the DFT study with a structure in which the HO_{RNA} was in the first coordination shell of Mg_A^{2+} (i.e., strongly coordinated—CHECK model; see Scheme 3). Such a model is

apparently more consistent with the two metal ion catalysis scheme, in which Mg_A^{2+} orients and activates the nucleophile by lowering its pK_a , facilitating the subsequent deprotonation.^{14,15} The closer the HO_{RNA} is to Mg_A^{2+} , the more easily this group will be deprotonated because the Mg_A^{2+} provides a greater stabilization to the product O^-_{RNA} . However, in the subsequent step of that mechanism, this very stable anionic product must unbind from Mg_A^{2+} to attack the substrate, and it is far from obvious that the energetic price for this step would be smaller than the advantage gained in the previous one.

In our DFT calculations, all attempts to attack the substrate starting from a structure with the HO_{RNA} strongly coordinated to Mg_A^{2+} have been unsuccessful. There is no stationary point with the HO_{RNA} deprotonated and bound to Mg_A^{2+} and its proton bound to any of the bases close by (Asp486 or O_α). Transfer of the proton to these bases leads to nonstationary points with energies of 17.0 and 16.8 kcal/mol above the unbound form. Moreover, the movement of the HO_{RNA} toward the substrate P_α causes the unbinding of the HO_{RNA} from Mg_A^{2+} , with a simultaneous decrease in energy, up to a minimum (stationary state) very similar to our DFT model structures in which the HO_{RNA} is weakly coordinated to Mg_A^{2+} . Therefore, the model rearranges and converts spontaneously into the weakly coordinated model, during the movement of the magnesium-bound HO_{RNA} toward the P_α (for nucleophilic attack).

We concluded that the structure with the HO_{RNA} bound to Mg_A^{2+} was outside the catalytic pathway, and therefore we concentrated subsequently on the structures that start with the HO_{RNA} weakly coordinated to Mg_A^{2+} . These will be discussed in detail.

The most straightforward hypothesis (HYP1, Scheme 2) is a proton transfer from HO_{RNA} to O_α . In this case, the deprotonation of the attacking nucleophile and protonation of the leaving group is achieved without the intervention of any additional group. The other three studied pathways (Scheme 2) correspond to the external HO^- hypothesis (HYP2), the transfer of the HO_{RNA} proton to an HO^- coordinated to Mg_A^{2+} (HYP3), and the transfer of the HO_{RNA} proton to the conserved protein residue Asp485 (HYP4). In all of these last three hypotheses, we found that protonation of the leaving group could be achieved by His1085. The external HO^- hypothesis resulted in the most feasible mechanism for RNAP II.

3.1. Intramolecular Proton Transfer Followed by Nucleophilic Attack (HYP1). This hypothesis of a mechanism begins with the transfer of the HO_{RNA} proton to the O_α atom. We have started with this case because it includes the two proton transfer reactions in a very straightforward way. The substrate O_α would act first as a base, deprotonating HO_{RNA} , and then act as an acid, protonating the leaving PP_i (substrate assisted catalysis). We have also tested the influence of the basis set used in the calculations in this specific reaction pathway.

The direct transfer of the HO_{RNA} proton to the O_α atom from the reactants (R) does not lead to a stationary point (the protonated O_α atom lies 25.1 kcal/mol above the reactants, Figure 2). However, upon rearrangement of the $H-O_{1\alpha}-P_\alpha-O_{5'}$ dihedral (such that this group engages in a hydrogen bond with an O_γ atom), we found a stationary state with the proton transferred to the O_α atom (structure I_{dRNA} in Figure 2). I_{dRNA} lies 24.9 kcal/mol above the initial reactants. The distance between ion O^-_{RNA} and atom P_α is 3.36 Å, and the distance between atom P_α and the ether oxygen atom of the scissile bond ($O_{3\beta}$) is 1.62 Å. We have not located the transition

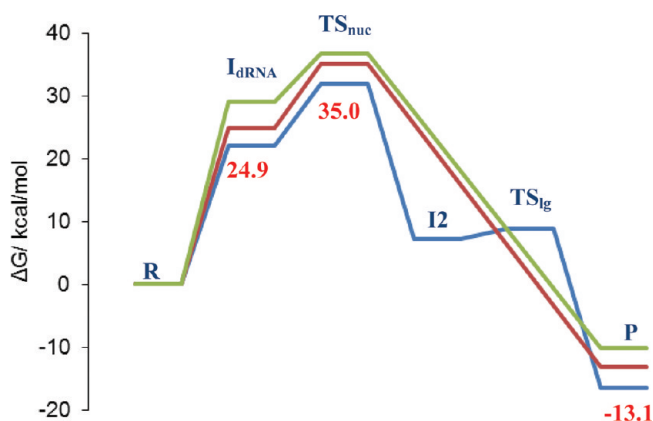


Figure 2. Potential energy surfaces for the HYP1 pathway calculated with different basis sets. Blue, B3LYP/6-31G(d)//ONIOM(B3LYP/6-31G(d):PM3MM); red, B3LYP/6-311++G(2d,2p)//ONIOM(B3LYP/6-31G(d):PM3MM); green, B3LYP/6-311++G(2d,2p)//ONIOM(B3LYP/6-311++G(2d,2p):PM3MM). As can be seen, the shape of the PES is the same whenever the optimization is done with the smaller or larger basis sets, as long as the final energies are calculated with the larger basis sets.

state between R and I_{dRNA} because the energy barrier for the subsequent steps of this pathway is too high to match the enzyme kinetics, as will be seen below. The transition state for the nucleophilic attack of ion O_{RNA}^- to the P_{α} atom (TS_{nuc} , Figure 3) is directly connected to I_{dRNA} and has an energy of 10.1 kcal/mol above that of the reactants (Figure 2). The bond between ion O_{RNA}^- and the P_{α} atom is being formed (2.14 Å), and the bond between the P_{α} and the $O_{3\beta}$ atoms is starting to break (1.73 Å). Decay from TS_{nuc} leads to structure I_2 , a pentacoordinated intermediate. In this structure, the $O_{RNA}^- \cdots P_{\alpha}$ distance is 1.75 Å and the $P_{\alpha} \cdots O_{3\beta}$ distance is 1.80 Å. I_2 is 13.1 kcal/mol above R. Elimination of PPi (the leaving group, lg) proceeds through a second transition state TS_{lg} , 11.5 kcal/mol above R. In TS_{lg} (Figure 3), the $O_{3\beta}$ atom is moving away from the P_{α} atom and starting to form a bond with the proton of the O_{α} atom. The $O_{RNA}^- \cdots P_{\alpha}$ distance is now 1.65 Å, and the $P_{\alpha} \cdots O_{3\beta}$ distance is 2.48 Å. Finally, the products (P) correspond to the protonated PPi and the substrate incorporated in the RNA chain. The products are 13.1 kcal/mol below R. The new phosphodiester bond ($O_{RNA}^- - P_{\alpha}$) has a bond length of 1.65 Å, and the $O_{3\beta}$ atom of PPi has moved 3.40 Å away from the P_{α} atom. Figure 3 shows a detail of the two transition states discussed in the text. To complete the catalytic cycle, the PPi must subsequently leave the active site, and RNA must translocate one base pair away from the active site. The reaction is exoenergetic, as expected, as we are breaking a highly energetic bond (Figures 2 and 3). The energy barrier between the reactants (R) and the highest point along the pathway (TS_{nuc}) amounts to 34.9 kcal/mol, which is clearly too high for an enzymatic reaction. The turnover number for the enzyme is 0.16 s^{-1} , from which a rate-limiting barrier of 18.1 kcal/mol can be derived. Consequently, regardless of whether the rate limiting step of the reaction is the conformational transition (trigger loop closing) or the chemical step, any of the associated barriers must lie below (or be equal to) 18.1 kcal/mol.

When dealing with negatively charged systems (such as this one), the use of diffuse functions is highly recommended. To be on the safe side we decided to reoptimize the structures using a

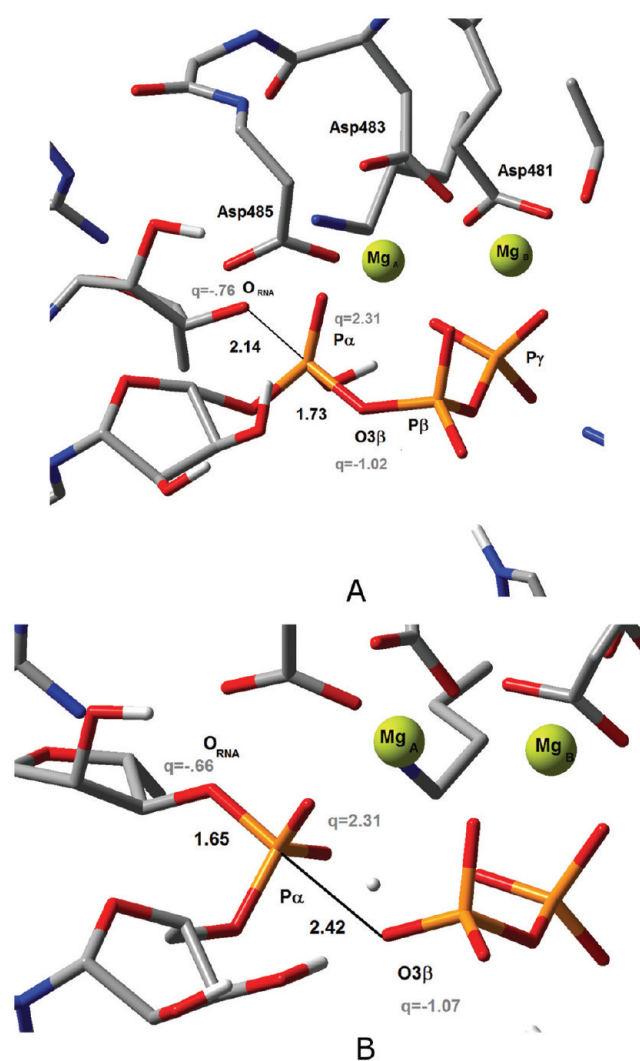


Figure 3. A detailed illustration of the structures of TS_{nuc} and TS_{lg} for HYP1. TS_{nuc} (A) is the transition state related with the nucleophilic attack of O_{RNA}^- to P_{α} . TS_{lg} (bottom) is the transition state associated with proton transfer to PPi and elimination of the PPi from the pentacoordinated complex. Relevant distances and Mulliken atomic charges q are shown.

triple- ζ basis set with diffuse functions and extra polarization functions, namely, 6-311G++(2d,2p) (Figure 2). A comparison of the three energy profiles shows that the B3LYP/6-311G++(2d,2p)//B3LYP/6-31G(d) curve represents a good compromise between PES accuracy and computational time. The smaller basis set identifies the pentacoordinated species as a transient short-lived intermediate. When the PES is calculated at the B3LYP/6-311G++(2d,2p)//B3LYP/6-31G(d) level, the energy of the second transition state becomes lower than that of the pentacoordinated intermediate, giving rise to a profile equivalent to the B3LYP/6-311G++(2d,2p)//B3LYP/6-311G++(2d,2p) PES.

As the activation and reaction energies obtained for the nucleophilic attack with this computational methodology were small (10.1 kcal/mol at the B3LYP/6-311G++(2d,2p)//B3LYP/6-31G(d) level), we were motivated to search for alternative residues that could deprotonate the HO_{RNA} more efficiently, eliminating the high barriers associated with this step.

3.2. The External HO⁻ Group hypothesis (HYP2). This hypothesis assumes that the proton of the HO_{RNA} group is abstracted by a HO⁻ ion coming from bulk solution. Therefore, one of the protein residues must protonate the PPI. The only residue that is appropriately positioned and has the adequate acidic character near the triphosphate of the substrate is His1085. In the reactants (R), we have included a HO⁻ ion establishing hydrogen bonds with both the 2'HO and the 3'HO of the RNA terminus. This position was chosen because it is the only stationary point we could find close enough from the HO_{RNA} group for an efficient deprotonation.

To measure the energy involved in transferring a HO⁻ ion from the bulk solvent to the active site, we needed to calculate the free energy of transfer and correct it for the nonstandard state associated with the low concentration of the ion at physiological pH. A preliminary search for the most stable and catalytic competent positions for the HO⁻ ion showed us that the correct place for the HO⁻ ion was doubly hydrogen bonded to both the hydroxyl groups of the terminal RNA, between them and the Mg²⁺ ion. This position is the most stable one that obeys the requisite of having the hydroxyl ion hydrogen bonded to the HO_{RNA} that will be deprotonated.

The free energy of transfer between the two media (bulk solvent and active site) was accounted for by a classical molecular dynamics thermodynamic integration and resulted in -4.20 ± 0.52 kcal/mol. The gain in water–water interactions and the strong interactions between the ion and the terminal RNA make this step favorable. However, in reality, the ion is present in the system at a much lower concentration of 10^{-7} mol·dm⁻³. The translational entropy lost in the confinement of the ion from its volume in bulk solution to a volume of 1–3 water molecules (the volume in which it can move without disturbing the two hydrogen bonds) is known to be 12.4–11.7 kcal/mol, respectively. This value is not very sensitive to the exact volume of confinement. Taking the volume of three water molecules as a reference, we end up with a cost of +7.5 kcal/mol for bringing a HO⁻ ion from bulk solution to the active site, hydrogen bonded to the HO_{RNA} group and ready for deprotonation.

In R, the hydrogen bonds have a length of 1.90 Å and 1.62 Å from the 2'HO and 3'HO (i.e., HO_{RNA}) groups. The distance between the HO_{RNA} group and the P_α atom is 3.48 Å. The barrier for deprotonation is so shallow that we were not able to optimize it. Each time we tried, the system decayed to I_{dirNA}, in which the HO_{RNA} proton is transferred to the HO⁻ group and the hydrogen bond with the 2'HO is broken. The distance between the O⁻_{RNA} ion and the new water molecule is 1.54 Å. The O⁻_{RNA} ion is now at 3.37 Å from the P_α atom. I_{dirNA} lies 4.9 kcal/mol below R. The proton transfer is energetically favorable (Figure 5). In the second step, the ε2 proton from the positively charged His1085 (H_ε⁺) is transferred to the O_{1β} atom of the triphosphate through the transition state TS_{lg} (Figure 4). The distance between the proton H_ε⁺ and the O_{1β} atom is 1.43 Å, and the His1085 N_ε···H_ε⁺ distance is 1.16 Å. The reaction barrier for this proton transfer is 7.2 kcal/mol. The order of the two proton transfers may be interchanged; i.e., the protonation of PPI may occur prior to deprotonation of the HO_{RNA} group. This will depend on the kinetics of diffusion of the HO⁻ ion into the active site. After the proton transfer (structure I_{lg}), the distance between O⁻_{RNA} and P_α is 3.06 Å. Subsequently, O⁻_{RNA} attacks P_α via a pentacoordinate transition state (TS_{nuc}) with an associated reaction barrier of 9.1 kcal/mol (Figure 4). In TS_{nuc}, the distance between O⁻_{RNA} and P_α is 2.17 Å, and the distance between the

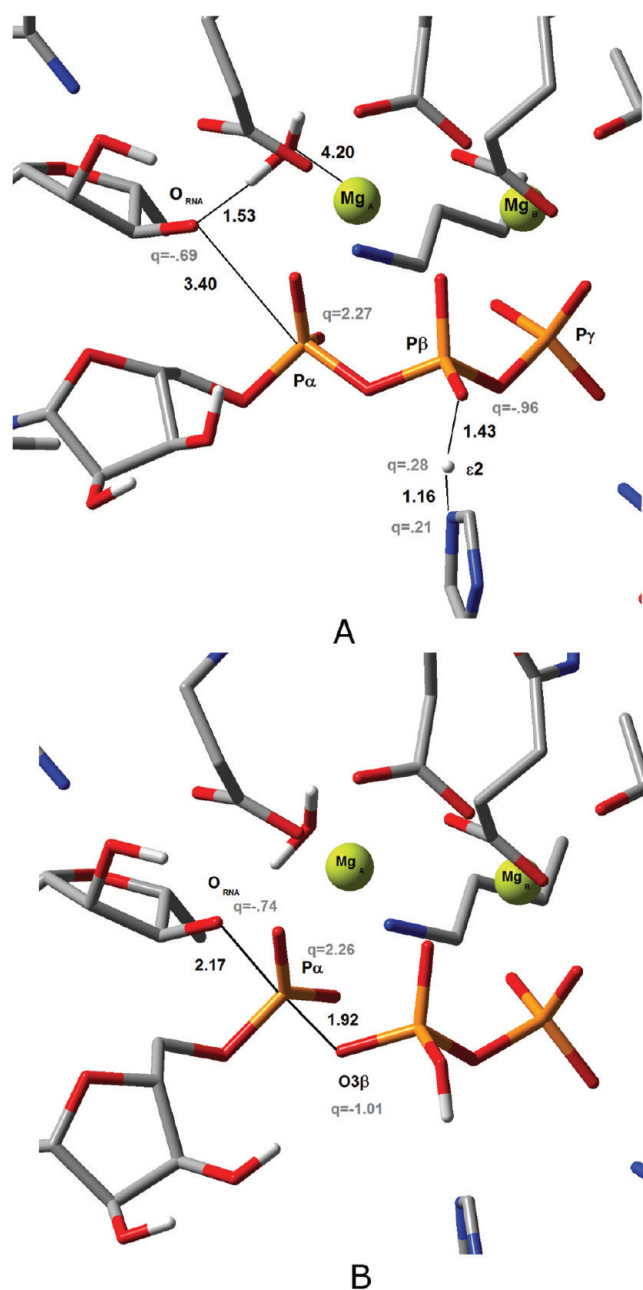


Figure 4. Structures of the transition states for hypothesis HYP2. TS_{lg} (top) is the transition state related with the proton transfer between His1085 and the O_β atom of the substrate triphosphate. TS_{nuc} (bottom) is the transition state associated with the nucleophilic attack of O⁻_{RNA} to P_α. Relevant distances and Mulliken atomic charges *q* are shown.

P_α and O_{3β} atoms amounts to 1.92 Å. This TS is very similar to the one for nucleophilic attack in HYP1. The energy of TS_{nuc} in relation to its reactants is 9.1 kcal/mol compared with the 10.5 kcal/mol of HYP1. The distance between O⁻_{RNA} and P_α is 2.14 Å in HYP1 and 2.17 in HYP2, and the distance between the P_α and O_{3β} atoms amounts to 1.73 Å in HYP1 and 1.92 Å in HYP2. Finally, in the product (P), the distance between the P_α and O_{3β} atoms increases to 3.70 Å, and the bond between P_α and O⁻_{RNA} is fully formed (1.68 Å). The product is 15.7 kcal/mol below TS_{nuc} and 13.3 kcal/mol below R, as expected, since we are breaking a highly energetic phosphodiester bond (Figure 5).

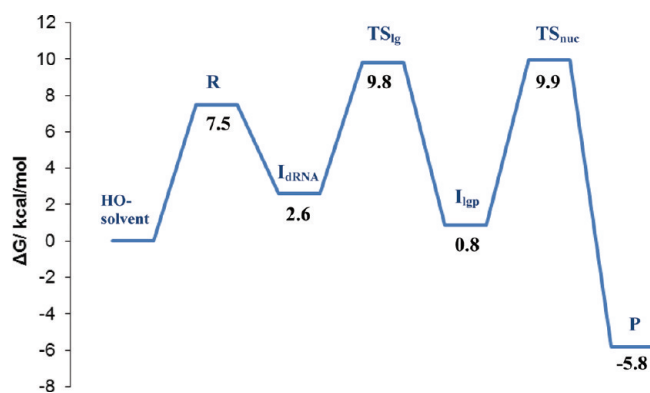


Figure 5. Potential energy surface for the mechanistic hypothesis HYP2. We have estimated that the transfer of the HO^- from the bulk solvent to the active site requires 7.5 kcal/mol. The proton transfer from His1085 (TS_{ig}) to the substrate triphosphate was a free energy barrier of 7.2 kcal/mol, and the nucleophilic attack (TS_{nuc}) was a free energy barrier of 9.1 kcal/mol. The products lie 13.3 kcal/mol below the reactants and 5.8 kcal/mol below the state with the hydroxyl in solution.

In total, starting from the hydroxyl ion in solution, the rate-limiting free energy barrier (the nucleophilic attack) amounts to 9.9 kcal/mol, and the reaction free energy corresponds to -5.8 kcal/mol.

3.3. Deprotonation by a HO^- Bound to Mg_A^{2+} and Protonation by His1085. A third hypothesis (HYP3) involves an initial proton transfer from the HO_{RNA} to a hydroxide ion bound to Mg_A^{2+} (instead of a free HO^- ion coming from the bulk solution) and the transfer of a proton from His1085 to P α , followed by nucleophilic attack. Even though the X-ray structure does not show such an ion coordinated to Mg_A^{2+} , this hypothesis shall not be excluded, as (i) the reduced water content of the structure can cause an impact on the existence of such ions and (ii) this hypothesis might correspond to an extension of the previous, in which the hydroxyl ion diffuses from the bulk solution and subsequently binds Mg_A^{2+} afterward. The difference here is that the ion would bind the Mg_A^{2+} before deprotonating the HO_{RNA} group.

The modeling of the HO^- ion in the coordination sphere of Mg_A^{2+} led to significant rearrangements. The $\text{O}_{2\beta}$ atom unbound from the Mg_A^{2+} cation, and the latter changed to a tetrahedral coordination geometry, getting closer to the HO_{RNA} group and away from Mg_B^{2+} .

The HO^- ion is now much closer to the substrate than in HYP2. Its negative charge polarizes further the substrate and increases the $\text{p}K_a$ of the $\text{O}_{1\beta}$ atom, which spontaneously abstracts a proton from His1085.

In the reactants (R_{ig}), the distance between the Mg^{2+} ions is 4.93 Å, the distance between the HO_{RNA} and Mg_A^{2+} is 3.67 Å, and the distance between Mg_A^{2+} and $\text{O}_{2\beta}$ is 4.14 Å. The 2'HO and 3'HO (HO_{RNA}) groups of the terminal RNA nucleotide both establish hydrogen bonds with the magnesium-bound HO^- ion, with distances of 1.72 Å and 1.61 Å. The distance between the HO_{RNA} and P α is 3.60 Å. The distance between HO^- and Mg_A^{2+} is 1.95 Å.

To see if the binding of HO^- to Mg_A^{2+} (HYP3) is thermodynamically favored over the weak interaction of the ion with the metal (HYP2), we disconnected the hydroxide from Mg_A^{2+} and reoptimized the structure, generating the reactants of HYP2. The energy dropped by 10.8 kcal/mol, showing that the structure of

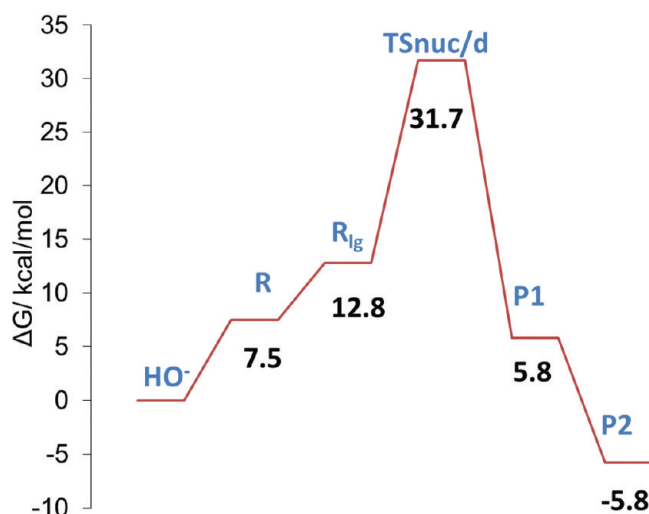
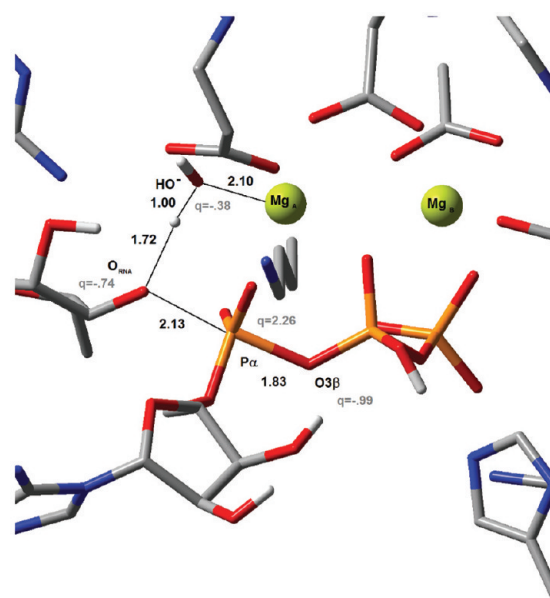


Figure 6. (Top) Structure of the transition state for HYP3 ($\text{TS}_{\text{nuc}/\text{dRNA}}$). The transition state is related to the nucleophilic attack of P α by O^-_{RNA} and the proton transfer from the HO_{RNA} group to the HO^- ion. Relevant distances and Mulliken atomic charges q are shown. (Bottom) Free energy profile for HYP3. The energy barrier for $\text{TS}_{\text{nuc}/\text{dRNA}}$ is 18.9 kcal/mol.

HYP2 (HO^- near but unbound from Mg_A^{2+}) is preferred. Anyway, the binding of HO^- to Mg_A^{2+} could lower the rate-limiting barrier for the reaction and bring an overall benefit in terms of the kinetics of the transformation. Therefore, we continued to follow this mechanism.

The transition state ($\text{TS}_{\text{nuc}/\text{dRNA}}$) corresponds to the nucleophilic attack on the P α atom, concerted with the transfer of the proton of the HO_{RNA} group to the metal bound HO^- ion (Figure 6). In this structure, the distance between the HO_{RNA} and P α groups shortens to 2.13 Å, and the distance between the P α and $\text{O}_{3\beta}$ atoms increases to 1.83 Å. The distance between the Mg_A^{2+} ion and the $\text{O}_{2\beta}$ atom changes to 2.28 Å, and Mg_A^{2+} becomes pentacoordinated again. The two Mg^{2+} ions approach

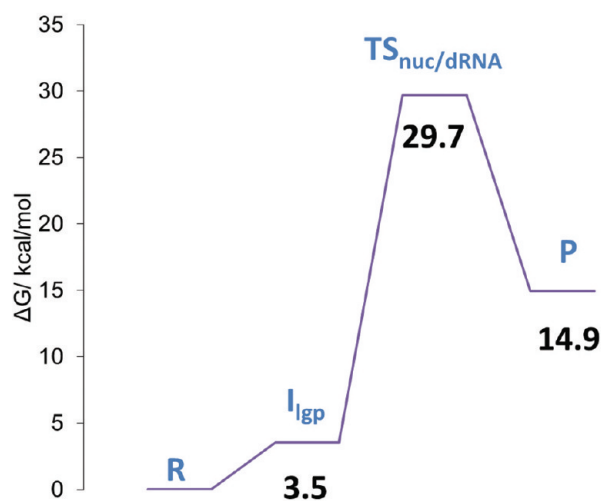
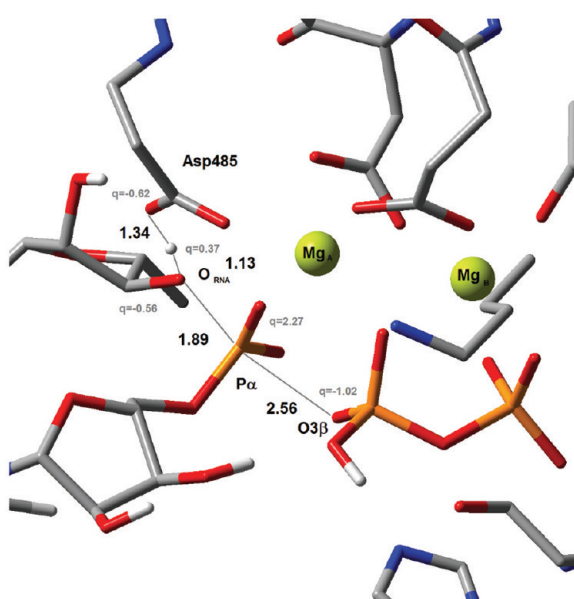


Figure 7. Structure of the transition state. The transition state is related with nucleophilic attack of HO_{RNA} to P_{α} and simultaneous proton transfer from the HO_{RNA} to Asp485 ($\text{TS}_{\text{nuc/dRNA}}$). Relevant distances and Mulliken atomic charges are shown. (Bottom) Free energy profile for HYP4. The free energy barrier for $\text{TS}_{\text{nuc/dRNA}}$ is 26.2 kcal/mol.

3.76 Å, which is closer to the values observed in HYP1 and HYP2. The energy barrier for this step is 18.9 kcal/mol. In this case, the nucleophilic attack is kinetically much less favorable than in HYP2 (Figure 6). In the products, the $\text{O}_{3\beta}$ atom separates from P_{α} (2.98 Å). The difference in energy between P and the reactants is -6.9 kcal/mol.

Comparing the final structures of HYP2 and HYP3, we can see that the difference lies in the coordination of the water molecule to Mg^{2+} , which is clearly unfavorable. Upon dissociation of the water molecule, the energy of the system drops by 11.6 kcal/mol, and the final reaction energy is -18.6 kcal/mol.

Starting from the initial system with the hydroxide in solution, the barrier amounts to 31.7 kcal/mol and the reaction energy to -5.8 kcal/mol.

Table 1. Reaction Steps for Each Hypothesis^a

	reaction steps
HYP1	<ol style="list-style-type: none"> 1 proton transfer from HO_{RNA} to O_{α} 2 nucleophilic attack ($\text{O}^{-}_{\text{RNA}}$ on P_{α})
HYP2	<ol style="list-style-type: none"> 1 HO^{-} transfer to active site 2 proton transfer HO_{RNA} to HO^{-} 3 proton transfer from His1085 to O_{β} 4 nucleophilic attack
HYP3	<ol style="list-style-type: none"> 1 HO^{-} transfer from bulk solvent to the active site 2 proton transfer from His1085 to O_{β} 3 nucleophilic attack and proton transfer from HO_{RNA} to HO^{-} 4 dissociation of H_2O from Mg_A^{2+}
HYP4	<ol style="list-style-type: none"> 1 proton transfer from His1085 to O_{β} 2 nucleophilic attack and proton transfer HO_{RNA} to Asp485

^a The CHECK hypothesis was not included because it does not lead to stationary points or it converts into other models already considered in the four hypotheses on the table.

3.4. Transfer of the Proton to Asp485 Concerted with Nucleophilic Attack. The last hypothesis (HYP4), frequently pointed out for DNA polymerases, is a proton transfer from the HO_{RNA} group to a conserved aspartate at the active site. In RNAP II, the only residue in position to fulfill this role is Asp485. However, when we tried to transfer the proton prior to the nucleophilic attack, our calculations have shown that the product of the proton transfer was not a stationary point in the potential energy surface and laid 33.6 kcal/mol above the reactants. Therefore, we tried a case in which the proton transfer from the HO_{RNA} group to Asp485 occurred simultaneously with the nucleophilic attack of the $\text{O}_{\text{RNA}}^{-}$ ion to the P_{α} atom (Figure 7). In the initial reactants (R), His1085 is protonated and the HO_{RNA} group is at 3.54 Å from the P_{α} . In I_{igp} , the proton is already in the triphosphate; the HO_{RNA} group is almost at the same distance from the P_{α} atom (3.57 Å). This structure is 3.5 kcal/mol above R. The structure is thus ready for the subsequent nucleophilic attack. From here, we found the transition state for that reaction ($\text{TS}_{\text{nuc/dRNA}}$), which in fact corresponded to the nucleophilic attack concerted with the proton transfer to Asp485. The $\text{HO}_{\text{RNA}}\cdots\text{HO}$ bond is elongated to 1.13 Å, and the distance between the HO_{RNA} proton and the Asp485 O_{γ} corresponds to 1.34 Å. The $\text{O}_{\text{RNA}}^{-}$ ion approached the P_{α} atom up to 1.89 Å, and the $\text{P}_{\alpha}\cdots\text{O}_{3\beta}$ distance elongated to 2.56 Å. The Mg_A^{2+} and Mg_B^{2+} ions came closer to each other (3.20 Å). The activation free energy for this step amounted to 26.2 kcal/mol (Figure 7). In the products (P), the bond between the P_{α} and the O_{β} atoms got broken (2.95 Å), and the new phosphodiester bond ($\text{O}^{-}_{\text{RNA}}-\text{P}_{\alpha}$) was formed (1.66 Å). The products' energies were 14.5 kcal/mol more than that of the reactants. The high activation energy excludes this hypothesis, as the pathway studied in HYP2 is kinetically much more favorable.

4. CONCLUSIONS

Multisubunit RNAPs are evolutionarily unrelated with both single RNAPs and DNAPs. Nevertheless, they catalyze the same

chemical reaction, the polymerization of nucleotides into a nucleic acid chain. The X-ray structures of RNAPII¹ have the Mg_A²⁺ ion in a different position than the one observed for some DNAPs, not located between the HO_{RNA} and P_α groups but instead above the P_α atom.

In the X-ray structures of polymerases with nucleic acid chains and the substrate, the nucleophile (the HO_{RNA/DNA}) is absent to prevent the conversion of the substrate into the exoenergetic product. Consequently, the arrangement of the HO_{RNA/DNA} in relation to the Mg_A²⁺ ion cannot be determined from the X-ray structures. The two-metal-ion scheme hypothesizes that the HO_{RNA} group should be coordinated to the Mg_A²⁺ ion at the beginning of the catalytic cycle. However, our DFT results strongly suggest that such a structure is less stable than a second one with the HO_{RNA} just weakly coordinated to the metal. This result is consistent with a molecular dynamics study of DNA polymerase β in which the coordination distance reported between the HO_{DNA} group and the Mg_A²⁺ ion in the reactant structure was 4.7 Å.²⁸ Moreover, if we start from a HO_{RNA} group strongly coordinated to the Mg_A²⁺ ion, the catalytic pathway necessarily evolves through a HO_{RNA}–Mg_A²⁺ unbinding before the proton transfers and the nucleophilic attack. The structure of the active site does not allow for a nucleophilic attack without a prior HO_{RNA}–Mg_A²⁺ dissociation. Further studies are needed to elucidate if the enzyme has the HO_{RNA} group bound to the Mg_A²⁺ ion before the catalytic cycle begins. However, what we can clearly conclude is that if such state exists it will represent only a precatalytic state and the dissociation of the HO_{RNA} group is a necessary condition for the catalytic reaction to begin.

We have studied the four possible hypotheses for the catalytic pathway. They mostly account for different participants in the acid/base catalyzed proton transfer reactions (deprotonation of HO_{RNA} and protonation of PPI) (Table 1). A comparison of the thermodynamic and kinetic profiles of the four pathways clearly shows that the one described in HYP2 (the external HO[−] group hypothesis) is by far the most favorable. This pathway involves deprotonation of the HO_{RNA} group by a bulk solvent hydroxide ion, protonation of PPI by His1085, and nucleophilic attack of the triphosphate by the O[−]_{RNA} ion. The rate-limiting step is the nucleophilic attack with an energy barrier of 9.9 kcal/mol relative to the initial state.

Our mechanism for RNAP II differs from the mechanisms proposed for DNAPs in the position of the HO_{RNA} nucleophile in relation to the Mg_A²⁺ ion. It also differs in the protonation of the leaving pyrophosphate, because we have taken into consideration the most recent findings by Castro et al. (role of His1085).^{2,3} The mechanism is consistent with the mechanisms of DNAPs by Flórian et al.^{10,29,30} in regard to the acceptors of the HO_{RNA/DNA} proton. The mechanism is consistent with the mechanisms for DNAPs by Alberts et al.²⁰ and Lin et al.⁵ in regard to the nature of the nucleophilic transition state. We found a single transition state associated with the nucleophilic attack with an O[−]_{RNA}···P_α distance of 2.17 Å (which compares to 2.20 in ref 5) and a P_α···O_{3β} distance of 1.92 Å (which compares to 1.90 in ref 5). Only one TS associated with the nucleophilic attack was found in both works.

■ ASSOCIATED CONTENT

Supporting Information. RMSDs for the α carbons. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: mjramos@fc.up.pt.

■ ACKNOWLEDGMENT

The authors thank the FCT—Fundação para a Ciência e Tecnologia—for financial support.

■ REFERENCES

- (1) Wang, D.; Bushnell, D. A.; Westover, K. D.; Kaplan, C. D.; Kornberg, R. D. Structural basis of transcription: role of the trigger loop in substrate specificity and catalysis. *Cell* **2006**, *127*, 941–54.
- (2) Castro, C.; Smidansky, E.; Maksimchuk, K. R.; Arnold, J. J.; Korneeva, V. S.; Gotte, M.; Konigsberg, W.; Cameron, C. E. Two proton transfers in the transition state for nucleotidyl transfer catalyzed by RNA- and DNA-dependent RNA and DNA polymerases. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 4267–4272.
- (3) Castro, C.; Smidansky, E. D.; Arnold, J. J.; Maksimchuk, K. R.; Moustafa, I.; Uchida, A.; Gotte, M.; Konigsberg, W.; Cameron, C. E. Nucleic acid polymerases use a general acid for nucleotidyl transfer. *Nat. Struct. Mol. Biol.* **2009**, *16*, 212–218.
- (4) Rittenhouse, R. C.; Apostoluk, W. K.; Miller, J. H.; Straatsma, T. P. Characterization of the active site of DNA polymerase beta by molecular dynamics and quantum chemical calculation. *Proteins* **2003**, *53*, 667–82.
- (5) Lin, P.; Pedersen, L. C.; Batra, V. K.; Beard, W. A.; Wilson, S. H.; Pedersen, L. G. Energy analysis of chemistry for correct insertion by DNA polymerase beta. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 13294–9.
- (6) Flórian, J.; Goodman, M. F.; Warshel, A. Computer simulation of the chemical catalysis of DNA polymerases: Discriminating between alternative nucleotide insertion mechanisms for T7 DNA polymerase. *J. Am. Chem. Soc.* **2003**, *125*, 8163–8177.
- (7) Abashkin, Y. G.; Erickson, J. W.; Burt, S. K. Quantum chemical investigation of enzymatic activity in DNA polymerase beta. A mechanistic study. *J. Phys. Chem. B* **2001**, *105*, 287–292.
- (8) Bojin, M. D.; Schlick, T. A quantum mechanical investigation of possible mechanisms for the nucleotidyl transfer reaction catalyzed by DNA polymerase beta. *J. Phys. Chem. B* **2007**, *111*, 11244–11252.
- (9) Florian, J.; Goodman, M. F.; Warshel, A. Computer simulation studies of the fidelity of DNA polymerases. *Biopolymers* **2003**, *68*, 286–299.
- (10) Florian, J.; Goodman, M. F.; Warshel, A. Computer simulations of protein functions: searching for the molecular origin of the replication fidelity of DNA polymerases. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 6819–24.
- (11) Fothergill, M.; Goodman, M. F.; Petruska, J.; Warshel, A. Structure-Energy Analysis of the Role of Metal-Ions in Phosphodiester Bond Hydrolysis by DNA-Polymerase-I. *J. Am. Chem. Soc.* **1995**, *117*, 11619–11627.
- (12) Kaplan, C. D.; Larsson, K. M.; Kornberg, R. D. The RNA polymerase II trigger loop functions in substrate selection and is directly targeted by alpha-amanitin. *Mol. Cell* **2008**, *30*, 547–556.
- (13) Beese, L. S.; Steitz, T. A. Structural Basis for the 3′-5′ Exonuclease Activity of Escherichia-Coli DNA-Polymerase-I - a 2 Metal-Ion Mechanism. *EMBO J.* **1991**, *10*, 25–33.
- (14) Steitz, T. A.; Steitz, J. A. A General 2-Metal-Ion Mechanism for Catalytic Rna. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 6498–6502.
- (15) Steitz, T. A. Structural biology - A mechanism for all polymerases. *Nature* **1998**, *391*, 231–232.
- (16) Yang, W.; Lee, J. Y.; Nowotny, M. Making and breaking nucleic acids: Two-Mg²⁺-ion catalysis and substrate specificity. *Mol. Cell* **2006**, *22*, 5–13.
- (17) Ribeiro, A. J. M.; Ramos, M. J.; Fernandes, P. A. Benchmarking of DFT Functionals for the Hydrolysis of Phosphodiester Bonds. *J. Chem. Theory Comput.* **2010**, *6*, 2281–2292.

(18) Noodleman, L.; Lovell, T.; Han, W.-G.; Li, J.; Himo, F. Quantum Chemical Studies of Intermediates and Reaction Pathways in Selected Enzymes and Catalytic Synthetic Systems. *Chem. Rev.* **2004**, *104*, 459–508.

(19) Leopoldini, M.; Marino, T.; Michelini, M.; Rivalta, I.; Russo, N.; Sicilia, E.; Toscano, M. The role of quantum chemistry in the elucidation of the elementary mechanisms of catalytic processes: from atoms, to surfaces, to enzymes. *Theor. Chem. Acc.* **2007**, *117*, 765–779.

(20) Alberts, I. L.; Wang, Y.; Schlick, T. DNA polymerase beta catalysis: Are different mechanisms possible? *J. Am. Chem. Soc.* **2007**, *129*, 11100–11110.

(21) Kluge, S.; Weston, J. Can a hydroxide ligand trigger a change in the coordination number of magnesium ions in biological systems? *Biochemistry* **2005**, *44*, 4877–85.

(22) Rodriguez-Cruz, S. E.; Jockusch, R. A.; Williams, E. R. Hydration energies and structures of alkaline earth metal ions, $M^{2+}(H_2O)_n$, $n=5-7$, $M = Mg, Ca, Sr$, and Ba . *J. Am. Chem. Soc.* **1999**, *121*, 8898–8906.

(23) Cisneros, G. A.; Perera, L.; Schaaper, R. M.; Pedersen, L. C.; London, R. E.; Pedersen, L. G.; Darden, T. A. Reaction Mechanism of the epsilon Subunit of E-coli DNA Polymerase III: Insights into Active Site Metal Coordination and Catalytically Significant Residues. *J. Am. Chem. Soc.* **2009**, *131*, 1550–1556.

(24) Focia, P. J.; Alam, H.; Lu, T.; Ramirez, U. D.; Freymann, D. M. Novel protein and Mg^{2+} configurations in the $Mg(2+)$ GDP complex of the SRP GTPase Ffh. *Protein: Struct. Funct. Genet.* **2004**, *54*, 222–230.

(25) Nowotny, M.; Yang, W. Stepwise analyses of metal ions in RNase H catalysis from substrate destabilization to product release. *EMBO J.* **2006**, *25*, 1924–1933.

(26) Kapustina, M.; Carter, C. W. Computational studies of tryptophanyl-tRNA synthetase: Activation of ATP by induced-fit. *J. Mol. Biol.* **2006**, *362*, 1159–1180.

(27) Kluge, S.; Weston, J. Can a hydroxide ligand trigger a change in the coordination number of magnesium ions in biological systems? *Biochemistry* **2005**, *44*, 4877–4885.

(28) Yang, L.; Arora, K.; Beard, W. A.; Wilson, S. H.; Schlick, T. Critical role of magnesium ions in DNA polymerase beta's closing and active site assembly. *J. Am. Chem. Soc.* **2004**, *126*, 8441–53.

(29) Florian, J.; Goodman, M. F.; Warshel, A. Computer simulation of the chemical catalysis of DNA polymerases: discriminating between alternative nucleotide insertion mechanisms for T7 DNA polymerase. *J. Am. Chem. Soc.* **2003**, *125*, 8163–77.

(30) Florian, J.; Goodman, M. F.; Warshel, A. Computer simulation studies of the fidelity of DNA polymerases. *Biopolymers* **2003**, *68*, 286–99.

Efficient Explicit-Solvent Molecular Dynamics Simulations of Molecular Association Kinetics: Methane/Methane, Na⁺/Cl⁻, Methane/Benzene, and K⁺/18-Crown-6 Ether

Matthew C. Zwier, Joseph W. Kaus, and Lillian T. Chong*

Department of Chemistry, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, United States

S Supporting Information

ABSTRACT: Atomically detailed views of molecular recognition events are of great interest to a variety of research areas in biology and chemistry. Here, we apply the weighted ensemble path sampling approach to improve the efficiency of explicit solvent molecular dynamics (MD) simulations in sampling molecular association events between two methane molecules, Na⁺ and Cl⁻ ions, methane and benzene, and the K⁺ ion and 18-crown-6 ether. Relative to brute force simulation, we obtain efficiency gains of at least 300 and 1100-fold for the most challenging system, K⁺/18-crown-6 ether, in terms of sampling the association rate constant *k* and distribution of times required to traverse transition paths, respectively. Our results indicate that weighted ensemble sampling is likely to allow for even greater efficiencies for more complex systems with higher barriers to molecular association.

1. INTRODUCTION

Proteins bind their partners in a highly specific manner. Understanding the mechanisms of these binding events is not only fundamentally interesting but could also impact fields such as protein engineering, host–guest chemistry, and drug discovery. Atomistic molecular dynamics (MD) simulations can potentially offer the most detailed views of molecular recognition events, especially when performed with explicit solvent. However, only up to a microsecond of simulation is practical on typical computing resources, while protein binding events require microseconds and beyond.¹ It is therefore computationally prohibitive to capture these events by sufficiently long “brute force” simulations. Fortunately, the long time scales required for protein binding events are not necessarily a result of the actual events taking a long time; instead, the events may be fast but infrequent, separated by long waiting times.

Path sampling approaches^{2–10} aim to capture rare events by minimizing the simulation of long waiting times between events.¹¹ Weighted ensemble sampling² is one such approach which is rigorously correct for any type of stochastic simulation,¹² easily parallelized, and simultaneously provides both transition paths and their associated kinetics.² Weighted ensemble sampling has been applied to Brownian dynamics simulations of protein–protein binding,² protein–substrate binding,¹³ protein folding,¹⁴ Monte Carlo simulations of large-scale conformational transitions in the molecular switches calmodulin¹⁵ and adenylate kinase,¹⁶ and molecular dynamics simulations of alanine dipeptide in implicit solvent.¹⁷

We apply the weighted ensemble path sampling approach with explicit-solvent MD simulations. Our goal is to determine the efficiency of the weighted ensemble approach relative to brute force simulation in sampling molecular associations for a range of well-studied systems: methane/methane,^{18–23} Na⁺/Cl⁻,^{24–33} methane/benzene,^{34,35} and K⁺/18-crown-6 ether^{36,37} (Figure 1). These systems were chosen because of their small size and

relatively low barriers to association ($\sim 2k_B T$); combined, these features make feasible the simulation of association events by brute force, providing us with opportunities to evaluate not only the efficiency of the weighted ensemble approach but its validity as well.

2. THEORY

2.1. Overview of Weighted Ensemble Sampling. Weighted ensemble sampling uses “statistical ratcheting” to efficiently sample rare events using stochastic simulations.^{2,11,15,17} To monitor the progress of these simulations toward the rare event of interest (here molecular association), a progress coordinate between the source (*A*, unbound) and destination (*B*, bound) states is defined by one or more order parameters; this progress coordinate is then divided into bins. A number of simulations are started in the unbound state *A*, which are then propagated for a fixed time τ . After this propagation time, if a simulation has progressed into a bin closer to the destination state *B*, its current state is used to start replicas of that simulation; these replicas diverge due to the stochastic nature of the underlying dynamics. Alternatively, if the simulation has regressed toward the source state *A*, it is effectively terminated. This resampling procedure¹² involving the replication of productive simulations and termination of unproductive simulations is repeated at fixed intervals (τ , 2τ , 3τ , and so on) until the desired number of rare events (crossings into state *B*) is sampled. Once a simulation reaches the destination state *B*, it is removed from the destination state *B* and “recycled” as a new simulation starting from the source state *A*. As this propagation and resampling procedure is repeated, the transition path ensemble—an ensemble of continuous trajectories between the source and destination states—is generated. As shown in Figure 2, some common history is shared among this

Received: November 2, 2010

Published: February 25, 2011

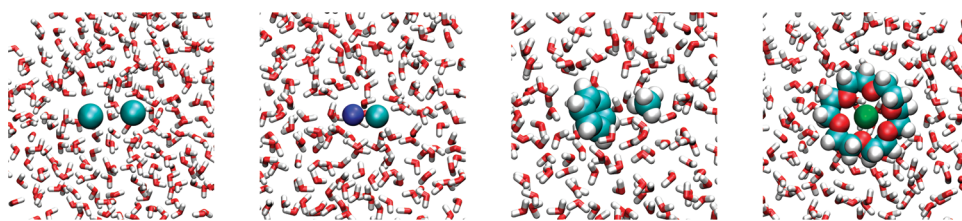


Figure 1. Molecular recognition systems of this study. From left to right, two methane molecules, Na^+/Cl^- , benzene and methane, and a K^+ ion with 18-crown-6 ether. All systems were immersed in explicit water molecules. (Prepared with VMD.³⁸)

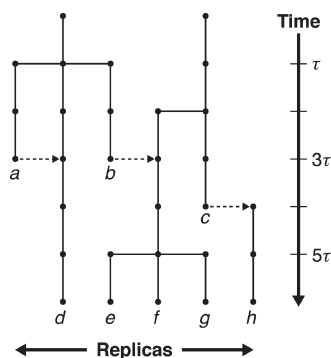


Figure 2. Schematic diagram of weighted ensemble molecular dynamics trajectories. Replication and termination of simulations occurs at intervals of τ , the propagation/resampling time. Trajectories *a* and *b* are terminated at $t = 3\tau$, and trajectory *c* reaches the destination state at $t = 4\tau$, at which time its statistical weight is assigned to a newly created replica which traces out trajectory *h*; the dotted arrows indicate a transfer of statistical weight but not history. Trajectories are replicated at $t = \tau$, $t = 2\tau$, and $t = 5\tau$. Note that trajectories *e*, *f*, and *g* share common history but are independent from trajectories *d* and *h*, which themselves are mutually independent.

ensemble of trajectories, and each trajectory has a maximum length τN_τ after N_τ iterations of propagation and resampling. When the trajectories are generated using molecular dynamics simulations, a stochastic thermostat is required to allow for divergence of trajectories after resampling.

To maintain correct statistics and kinetics of the transition paths, each simulation is assigned an appropriate statistical weight. When simulations are replicated, their statistical weights are split. When simulations are terminated for regressing toward the source state *A*, their statistical weights are merged into existing replicas, and when simulations are terminated for reaching the destination state *B*, their statistical weights are removed from the destination state *B* and assigned to newly created replicas in the initial state *A*.

2.2. Rate Constants. Weighted ensemble sampling yields kinetic information as a simulation progresses. After steady-state probability recycling is attained, the rate constant k is given by the average probability current I_B into the destination state *B*:^{2,15,39}

$$k = \langle I_B \rangle \quad (1)$$

where the angle brackets indicate a time average. Because the recycling procedure described above eliminates all probability from the destination state *B* at each resampling, the probability current I_B may be approximated as

$$I_B \approx \frac{P_B(\tau N_\tau)}{\tau} \quad (2)$$

where τ is the weighted ensemble propagation/resampling time step and $P_B(\tau N_\tau)$ is probability contained in the destination state at time τN_τ (weighted ensemble iteration N_τ) immediately prior to recycling. Since $P_B(\tau N_\tau)$ must be monitored in order to ensure probability conservation during a weighted ensemble simulation, the rate constant k is obtained “for free.”

The partially shared history of weighted ensemble trajectories results in highly correlated probability current measurements; that is, $I_B(\tau N_\tau)$ and $I_B(\tau[N_\tau - 1])$ are not statistically independent. The time average $\langle I_B \rangle$ may be computed in the usual way, but the associated confidence interval (encompassing the statistical error in the rate constant) must be computed with a method that accounts for the time correlation within I_B , such as Monte Carlo bootstrapping.^{2,40,41}

On the other hand, the quantity accessible from brute force dynamics is not the probability current into the destination state but is rather a set of elapsed times between completed transition events. That is, brute force simulation does not yield the rate constant directly, but rather the first passage time distribution. For transitions dominated by a single time scale, this distribution is exponential, and the rate constant is simply the inverse of the mean first passage time $\langle t_{\text{fp}} \rangle$:³⁹

$$k = \langle t_{\text{fp}} \rangle^{-1} \quad (3)$$

It should be noted that these two methods for determining the rate constant k are alternative mathematical descriptions of the same underlying physical principles (for an extensive discussion, see ref 39). Thus, rate constants obtained from brute force and weighted ensemble simulations may be directly compared, given that the same model was used for propagating dynamics in both cases and that the confidence interval for the rate constant is calculated correctly for the weighted ensemble simulation.

2.3. Transition Event Durations. If a system exhibits rare but fast events, then the transition event duration t_{ed} —the amount of time it takes a transition to complete once it starts—is much less than the mean first passage time $\langle t_{\text{fp}} \rangle$ (which includes the waiting time between rare events):

$$t_{\text{ed}} \ll \langle t_{\text{fp}} \rangle$$

The probability distribution of t_{ed} , $F(t_{\text{ed}})$, is at least approximately an indicator of the extent of sampling of the transition pathways. Distinct pathways will have associated characteristic transition durations,⁴² and as transition pathways are sampled, $F(t_{\text{ed}})$ will become better resolved. Thus, the self-convergence of $F(t_{\text{ed}})$ is a strong indicator that the transition path ensemble has been adequately explored.

The transition event duration distribution $F(t_{\text{ed}})$ is built up directly from simulation trajectories simply by noting the time elapsed between exiting the source state *A* and entering the destination state *B*. In the brute force case, a set of event

durations is transformed directly into a cumulative distribution function in the usual manner (by counting the number of t_{ed} less than a specified value):

$$F(t_{\text{ed}}) = \frac{1}{N} \sum_i h(t_{\text{ed}(i)}) \quad (4)$$

where i indexes transitions, N is the number of transition events observed, $t_{\text{ed}(i)}$ is the duration of transition event i , and h is an indicator function satisfying

$$h(t_{\text{ed}(i)}) = \begin{cases} 1 & \text{if } t_{\text{ed}(i)} \leq t_{\text{ed}} \\ 0 & \text{otherwise} \end{cases}$$

Weighted ensemble simulations, on the other hand, yield not the set of event durations $\{t_{\text{ed}}\}$ but a set $\{(t_{\text{ed}}, w)\}$ of transition event durations and corresponding terminal statistical weights. These terminal weights partially encode the probability of arriving at the final state, and so a weighted variation of eq 4 must be used:

$$F(t_{\text{ed}}) = \frac{\sum_i w_i h(t_{\text{ed}(i)})}{\sum_i w_i} \quad (5)$$

There are several advantages to describing the transition event duration distribution as an (empirical) cumulative distribution function. First, rigorous confidence bands may be assigned to empirical distribution functions,^{43,44} allowing one to assign error bars to the entire t_{ed} distribution and facilitating the comparison of simulation results. Second, the number of points N_e in a realization of $F(t_{\text{ed}})$ is equal to the number of unique transition event durations sampled and, as such, can be considered a statistical sample size for the purposes of quantifying sampling, even in the weighted ensemble case. For this same reason, eq 5, despite being cast in a weighted form, describes a formal empirical distribution function and is therefore an unbiased estimator of the true cumulative distribution function.⁴⁴

2.4. Relative Efficiency of Weighted Ensemble Simulations. Any meaningful metric for comparing the relative efficiencies of weighted ensemble and brute force simulations must account for not only the computational expense of obtaining an estimate on a quantity such as the rate constant, but also the uncertainty of that estimate. In other words, an efficiency metric must take error bars into account. For a given quantity like the reaction rate k , we define the efficiency of weighted ensemble sampling relative to brute force as

$$S = \frac{t_{\text{(WE)}}}{t_{\text{eff}}} \quad (6)$$

where $t_{\text{(WE)}}$ is the aggregate weighted ensemble simulation time (not overcounting shared history) and t_{eff} is the effective amount of brute force simulation time that would be required to obtain an estimate with the same size error bar as that obtained from a weighted ensemble simulation. Consideration of the error structure of brute force simulations and application of eq 6 gives the following efficiency metrics S_k and S_{ed} for sampling of the association rate constant k and t_{ed} distribution, respectively:

$$S_k = \frac{t_{\text{(BF)}}}{t_{\text{(WE)}} \left(\frac{\Delta k_{\text{(BF)}}^*}{\Delta k_{\text{(WE)}}^*} \right)^2} \quad (7)$$

$$S_{\text{ed}} = \frac{t_{\text{(BF)}}}{t_{\text{(WE)}} \left(\frac{N_{e(\text{WE})}}{N_{e(\text{BF})}} \right)} \quad (8)$$

where t represents total simulation time, Δk^* is the width of the 95% confidence interval on the rate constant k relative to the time average $\langle k \rangle$, and N_e is the number of unique time values in the empirical distribution function $F(t_{\text{ed}})$; the subscripts (BF) and (WE) represent values from brute force and weighted ensemble simulations, respectively. Detailed derivations of eqs 7 and 8 are provided in the Supporting Information.

3. METHODS

3.1. Model Systems. Four systems were used to test the feasibility of using weighted ensemble sampling with explicit-solvent MD simulations to study molecular association events. These systems all possess simple, one-dimensional progress coordinates by which it is possible to unambiguously define “how close to binding” a simulation is at any point in time. All systems were immersed in boxes of explicit water molecules. The model systems in order of progressively more challenging features are described below.

Methane/Methane. This system is a simple example of a hydrophobic interaction. The natural progress coordinate of this system is simply the center-to-center distance between the two methane molecules.

Na⁺/Cl⁻. This system is a simple example of an electrostatic interaction. The natural progress coordinate of this system is the center-to-center distance between the two ions.

Methane/Benzene. Like the methane/methane system, methane/benzene is a model of hydrophobic interactions, but unlike the previous two systems, it does not exhibit an effective spherical symmetry. However, our brute force simulations of this system revealed that the condensed-phase bound state involves precession of the methane molecule about the surface of the benzene ring. Therefore, despite the broken spherical symmetry, the natural progress coordinate for this system is effectively one-dimensional and was taken to be the distance between the methane carbon and the center of mass of the benzene carbon atoms.

K⁺/18-crown-6 ether. This system is a simple example of the binding of a (trivially) rigid substrate (K⁺) by a flexible partner (18-crown-6 ether). Like methane/benzene, this system does not exhibit effective spherical symmetry. However, both simulation^{36,37} and X-ray crystallography⁴⁵ have indicated that the bound structure consists of the K⁺ ion coplanar with the crown ether oxygen atoms. The natural progress coordinate for this system is therefore the distance between the K⁺ ion and the center of mass of the ether oxygen atoms.

3.2. Simulation Details. Both brute force and weighted ensemble simulations were performed using the GROMACS 4.0.5 software package.⁴⁶ Production dynamics (both brute force and weighted ensemble) were propagated in the canonical (NVT) ensemble at 300 K using a Langevin thermostat⁴⁷ (coupling time 1 ps). Van der Waals interactions were switched off smoothly between 8 and 9 Å; to account for the truncation of the van der Waals interactions, a long-range analytical dispersion correction⁴⁸ was applied to energy and pressure. Real-space electrostatic interactions were truncated at 10 Å. Long range electrostatic interactions were calculated using particle mesh

Ewald⁴⁹ (PME) summation. Bonds to hydrogen atoms were constrained to their equilibrium lengths using LINCS,⁵⁰ permitting a 2 fs integration time step.

Each model system was constructed in its unbound state and solvated in a dodecahedral periodic box with a minimum 12 Å clearance between the solutes and the box walls. Following a 1000-step steepest-descent energy minimization, each system was subjected to 20 ps of NVT thermal equilibration followed by 1 ns of constant-pressure (NPT) density equilibration using a weak isotropic Berendsen barostat⁵¹ (reference pressure 1 bar, coupling time 5 ps, and compressibility $4.5 \times 10^{-5} \text{ bar}^{-1}$). In both equilibration stages, all heavy atoms were restrained using a harmonic potential. The resulting equilibrated systems were used as starting points for both brute force and weighted ensemble MD simulations. The initial pair separations were 10, 10, 17, and 15 Å for methane/methane, Na^+/Cl^- , methane/benzene, and $\text{K}^+/\text{18-crown-6}$ ether, respectively. The GROMOS 45A3 united-atom force field⁵² and SPC/E⁵³ water model were used for methane/methane and Na^+/Cl^- , while the OPLS-AA/L force field⁵⁴ and the TIP3P⁵⁵ water model were used for methane/benzene and $\text{K}^+/\text{18-crown-6}$ ether. Atom type assignments for $\text{K}^+/\text{18-crown-6}$ ether ether are provided in the Supporting Information (Figure S1).

3.3. Brute Force Dynamics Propagation. Brute force simulations for all model systems were started from the end points of their respective second-stage (density) equilibration runs. Each simulation was continued until a sufficient number of transition events was observed, with solute positions recorded every 10 fs. The methane/methane and methane/benzene systems were both run as single 1 μs trajectories. Na^+/Cl^- and $\text{K}^+/\text{18-crown-6}$ ether required multiple independent trajectories to observe a sufficient number of transition events. A total of 10 independent 1 μs trajectories were run for Na^+/Cl^- , and 100 independent 100 ns trajectories were run for $\text{K}^+/\text{18-crown-6}$ ether.

3.4. Determination of Bound and Unbound States. The analysis of brute force trajectories and the construction of weighted ensemble simulations require unambiguous definitions of bound and unbound states for each system. Because all four model systems possess one-dimensional progress coordinates, the same protocol for determining these states was applied to all four model systems. Pairwise condensed-phase interactions can be described by the potential of mean force (PMF) $u(r)$, the free energy of the system as a function of pair separation r .⁵⁶ Taking the zero of energy to be the noninteracting limit, for constant-volume systems the $u(r)$ is given by the following:²³

$$u(r)/k_{\text{B}}T = - \left(\ln \frac{P(r)}{r^2} - \ln \frac{P(r_0)}{r_0^2} \right) \quad (9)$$

where $P(r)$ is the probability of observing the system at a pair separation r , r_0 is the shortest distance at which the pair is effectively noninteracting ($du/dr \approx 0$ for all $r > r_0$), and the factors of r^2 arise from the transformation between the Cartesian coordinates of the MD simulation and the spherical polar coordinates in which $u(r)$ is expressed. For each model system, the PMF $u(r)$ was determined using eq 9 with pairwise distance probabilities $P(r)$ taken from the brute force trajectories. The unbound state A was defined as $A = \{r: r \geq r_0\}$, where (as above) r_0 is the shortest distance at which the pair is effectively noninteracting. This definition ensures that binding events

observed in brute force simulations are very nearly statistically independent. The bound state B was readily identified as being near the global minimum of $u(r)$ and defined as $B = \{r: r < r_B\}$, where r_B is the separation at which the global minimum well of $u(r)$ becomes concave up; that is, B is the basin of attraction of the global minimum of $u(r)$. The remainder of progress coordinate space defines a transition region $T = \{r: r_B \leq r < r_0\}$ wherein the partners are interacting but not definitively bound. PMF curves for each system are provided in Figures S3–S6 of the Supporting Information.

3.5. Determination of Weighted Ensemble Simulation Parameters. In addition to definitions of bound and unbound states, a weighted ensemble simulation requires selection of optimal bin sizes, numbers of replicas per bin, and propagation/resampling interval τ . In making these selections, the extent of sampling should be maximized (generally meaning more bins and more replicas per bin) while minimizing the overall computational cost (generally meaning fewer bins and fewer replicas per bin).

For all four model systems, the potential of mean force was used to determine a bin spacing aimed at maximizing the “ratcheting” effect of the weighted ensemble approach. Where the PMF was changing rapidly with respect to pair separation, bin boundaries were chosen such that the crossing of a bin does not require climbing more than $\sim k_{\text{B}}T$ in energy as indicated by the appropriate PMF. This ensures that the system can move about the progress coordinate with relative ease. Conversely, in the region where the PMF is slowly varying, a constant spacing of bins was adopted. The propagation period τ was then chosen so that the RMS change in pair separation over a time τ was approximately equal to the width of the bins in the slowly varying region of the PMF. This resulted in bins of width ~ 0.1 – 1.0 Å. Initial tests indicated that 50 replicas per bin yielded sufficiently precise values for the rate constant k at a reasonable computational cost, so this value was used for all four model systems. Detailed listings of the resulting bin boundaries are provided in Figures S3–S6 (Supporting Information), and the remaining weighted ensemble sampling parameters are summarized in Table S1 (Supporting Information).

3.6. Weighted Ensemble Dynamics Propagation. Weighted ensemble dynamics runs used exactly the same simulation parameters (force field, thermostat parameters, box volume, etc.) as those of the corresponding brute force simulations. As with the brute force simulations, the initial atomic coordinates and velocities were taken from the end of the equilibration phase for each model system. The weighted ensemble sampling algorithm was implemented in an in-house computer code as described above. Replicas were propagated in parallel on 32–96 CPU cores, requiring a few days to simulate each model system. Both the rate constant k and the transition event duration distribution $F(t_{\text{ed}})$ were monitored every 50 or 100 τ , and the weighted ensemble simulation was terminated when k was constant within uncertainty and $F(t_{\text{ed}})$ had converged to within 95% confidence and remained at that level, as determined by a two-sided Kolmogorov–Smirnov test⁴⁴ (a standard test of the statistical equivalence of two empirical distribution functions). Though resampling was performed with a period of τ , all analysis of the simulations was conducted at a time resolution of 10 fs (the period with which solute positions were recorded during the underlying dynamics simulations). The resulting aggregate simulation times for each system are presented in Table S2 (Supporting Information).

Table 1. Brute Force (BF) and Weighted Ensemble (WE) Aggregate Simulation Times t , Rate Constants (k), and Relative Sampling Efficiencies (S_k) for the Four Model Systems^a

system	t_{BF}	t_{WE}	k_{BF} (ps ⁻¹)	k_{WE} (ps ⁻¹)	S_k
methane/methane	1 μ s	299 ns	$1.91 \pm 0.10 \times 10^{-3}$	$1.61 \pm 0.06 \times 10^{-3}$	7.0
Na ⁺ /Cl ⁻	10 μ s	3.86 μ s	$1.86 \pm 0.09 \times 10^{-4}$	$1.82 \pm 0.11 \times 10^{-4}$	1.4
methane/benzene	1 μ s	369 ns	$8.6 \pm 0.7 \times 10^{-4}$	$7.7 \pm 0.3 \times 10^{-4}$	8.7
K ⁺ /18-crown-6	10 μ s	322 ns	$2.1 \pm 0.3 \times 10^{-5}$	$4.8 \pm 0.2 \times 10^{-5}$	300

^aAggregate simulation times correspond to the combined length of all trajectories (either brute force or weighted ensemble) for each system, without overcounting common history in the case of weighted ensemble simulations. Uncertainties on the rate constants represent 95% confidence intervals. Relative efficiencies were calculated using eq 7.

Table 2. Ratios of Rate Constants k and Average Waiting Times $\langle t_w \rangle$ for Brute Force (BF) and Weighted Ensemble (WE) Simulations

system	$k_{\text{(WE)}}/k_{\text{(BF)}}$	$\langle t_w \rangle_{\text{(BF)}}/\langle t_w \rangle_{\text{(WE)}}$
methane/methane	0.842	0.841
Na ⁺ /Cl ⁻	0.977	0.977
methane/benzene	0.827	0.822
K ⁺ /18-crown-6	1.93	1.94

4. RESULTS AND DISCUSSION

The purpose of this study was to determine the efficiency of weighted ensemble sampling relative to brute force sampling for association events in four molecular recognition systems. As described above, both the association rate constant k and the transition event duration distribution $F(t_{\text{ed}})$ can be used to quantify sampling of the transition path ensemble. We compare the efficiency and accuracy of weighted ensemble simulations relative to brute force simulations in terms of both rate constants and transition event distributions.

4.1. Rate Constants. The rate constant (k) values for brute force and weighted ensemble simulations were separately converged to within statistical uncertainty. As shown in Table 1, the weighted ensemble simulations are in qualitative agreement with brute force simulations for all systems; quantitative agreement was achieved for Na⁺/Cl⁻ and methane/benzene. The relative efficiency S_k of weighted ensemble sampling of the rate constant was modest (1.4-fold) for Na⁺/Cl⁻, greater than 5-fold for the diffusive systems (methane/methane and methane/benzene), and 300-fold for the most complex system, K⁺/18-crown-6 ether.

It is not surprising that the rate constant obtained by weighted ensemble sampling for K⁺/18-crown-6 ether does not agree with the brute force simulation, as the brute force $F(t_{\text{ed}})$ did not converge; it is less clear why the rate constants for methane/methane are not in agreement. One possibility is that either the brute force or the weighted ensemble simulation did not sample the full set of waiting times between rare events. The waiting time t_w between subsequent A \rightarrow B transition events relates the first passage time t_{fp} and the transition event duration t_{ed} according to

$$t_{\text{fp}} = t_{\text{ed}} + t_w$$

In all cases (including that in which t_{ed} and t_w are not statistically independent):

$$\langle t_{\text{fp}} \rangle = \langle t_{\text{ed}} \rangle + \langle t_w \rangle$$

where the angle brackets denote the expectation (mean) value. Since $\langle t_{\text{ed}} \rangle \ll \langle t_{\text{fp}} \rangle$ for all four systems considered here, the discrepancy between brute force and weighted ensemble

simulations in mean waiting time $\langle t_w \rangle$ accounts almost completely for the discrepancy in rate constants between simulation techniques (see Table 2). It is likely that the overestimated brute force waiting time for K⁺/18-crown-6 ether is due to poor convergence of the brute force simulation. Similarly, it seems likely that the methane/methane brute force simulation underestimated t_w for that system. In both of these cases, the efficiencies presented in Table 1 represent lower bounds, as they assume complete convergence of the brute force simulations.

Implicit in the foregoing analysis is the assumption that the first passage time distribution $F(t_{\text{fp}})$ obtained from the brute force simulation is exponential, as would be the case in a system possessing (effectively) a single barrier of constant height; that is,

$$F(t_{\text{fp}}) = 1 - \exp(-kt_{\text{fp}}) \quad (10)$$

where k is the rate constant. In this case, the rate constant k is equal to the inverse mean first passage time [cf. eq 3]. If the first passage time distribution $F(t_{\text{fp}})$ is not exponential, then the inverse mean first passage time is at best an approximation of the true rate constant; conversely, the weighted ensemble approach samples k directly, and so it can be expected to recover the correct rate constant (within the bounds of statistical uncertainty) regardless of whether the underlying physical mechanisms lead to an exponential first passage time distribution. For three of the four model systems (Na⁺/Cl⁻, methane/benzene, and K⁺/18-crown-6), the first passage time distributions obtained from brute force simulations conform to eq 10 to within 95% confidence (see Figure S2, Supporting Information). For methane/methane, however, the first passage time distribution deviates from the expected exponential distribution for $t_{\text{fp}} \lesssim 300$ ps. This offers an alternative explanation for why the rate constant values obtained for methane/methane differ between brute force and weighted ensemble simulations: because the first passage time distribution $F(t_{\text{fp}})$ is not exponential, the rate constant k obtained from the brute force first passage time distribution as $\langle t_{\text{fp}} \rangle^{-1}$ may in fact be inaccurate.

4.2. Transition Event Duration Distributions. In general, the weighted ensemble simulations were as good or better than brute force simulations in generating well-resolved transition event duration distributions $F(t_{\text{ed}})$. As shown in Figure 3, $F(t_{\text{ed}})$ was well-resolved by both brute force and weighted ensemble simulations for all systems except K⁺/18-crown-6 ether, for which brute force sampling was not capable of providing a converged $F(t_{\text{ed}})$ distribution. The resolution of distributions from weighted ensemble simulations far exceeds that of distributions obtained from brute force simulations, as demonstrated in the increased number N_c of unique transition durations sampled (see Table 3). Further, pathways generated by weighted ensemble sampling and having different transition event durations were

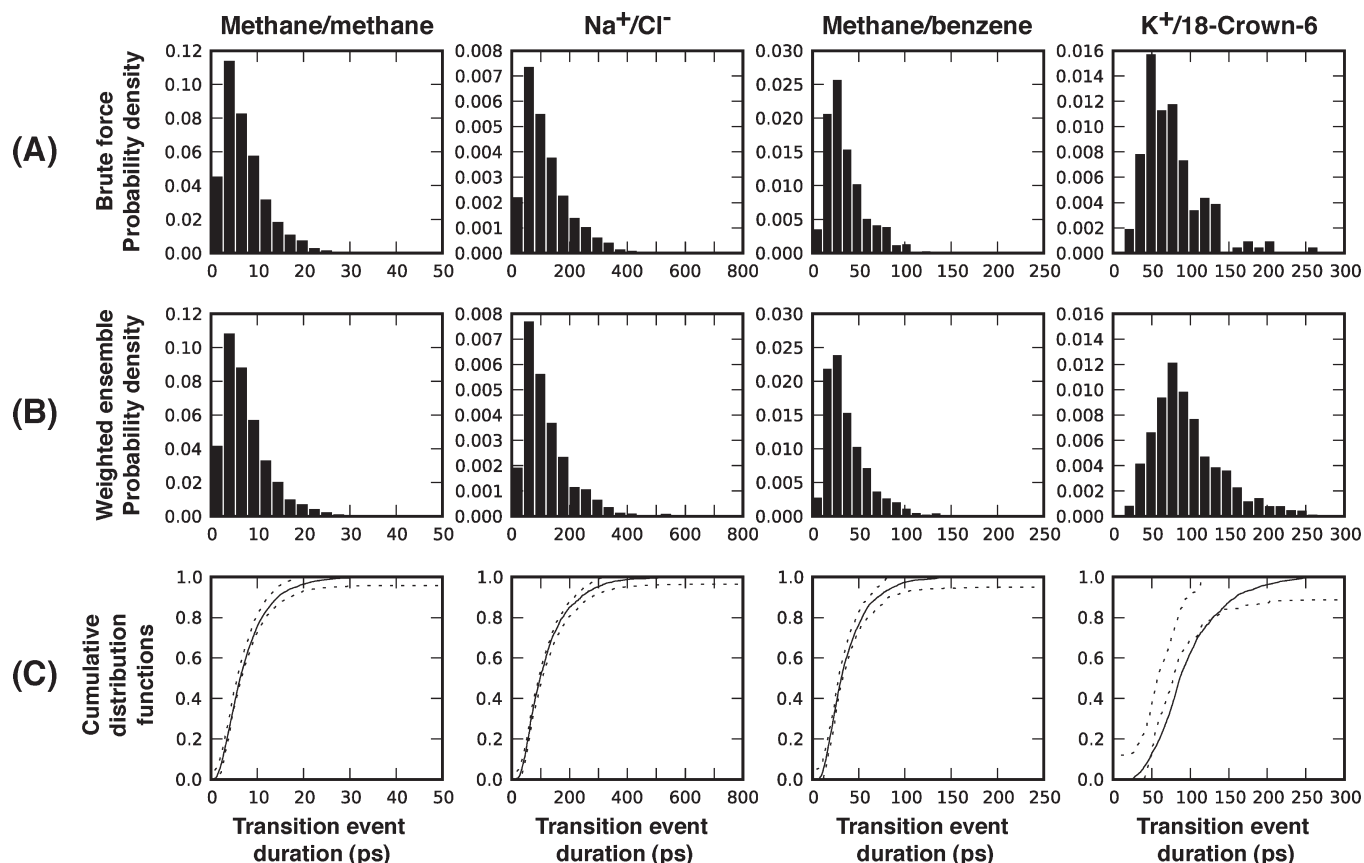


Figure 3. Transition event duration distributions obtained from (A) brute force and (B) weighted ensemble simulations. The cumulative distribution function (CDF) of the transition event duration probability for each model system is shown in (C). The brute-force CDF is plotted as a 95% confidence interval with dotted lines, and the solid line is the CDF obtained from the weighted ensemble simulation.

Table 3. Number of Unique Transition Durations N_e and Relative Efficiency S_{ed} of Sampling of the Transition Event Duration Distribution for Brute Force (BF) and Weighted Ensemble (WE) Simulations^a

system	$N_{e(BF)}$	$N_{e(WE)}$	S_{ed}
methane/methane	1021	2304	7.5
Na^+/Cl^-	1415	8780	16
methane/benzene	750	5485	20
$\text{K}^+/\text{18-crown-6}$	145	5007	1100

^a Relative efficiency was calculated using eq 8.

indeed noticeably different from each other (see Supporting Information, section S.1, Figures S1 and S2, and Movies S1 and S2). These are strong indications that the weighted ensemble algorithm effectively enhances sampling of the transition path ensemble. The relative efficiency S_{ed} of sampling $F(t_{ed})$ increased with the complexity of the molecular recognition system, ranging from 1 to 3 orders of magnitude. The 1100-fold relative efficiency of weighted ensemble sampling for $\text{K}^+/\text{18-crown-6}$ ether is a conservative estimate, as the referenced brute force simulation had not even reached convergence with respect to $F(t_{ed})$.

As shown in Tables 1 and 3, $S_k < S_{ed}$ in all four cases. This is partly a consequence of our definitions of the efficiency metrics S_k and S_{ed} (see above and Supporting Information), but it also reflects that the rate constant k is generally more difficult to sample than the set of transition event durations $\{t_{ed}\}$. In

particular, convergence of the rate constant k requires sampling of *all* important pathways as well as a steady state flow of probability through them.

4.3. How Much Sampling Is Required? As evident for $\text{K}^+/\text{18-crown-6}$ ether, the most complex system of this study, it is not always possible to obtain converged brute force simulations of molecular association events. In such cases, how does one know if the weighted ensemble approach has achieved sufficient sampling? One can, at least, gauge the self-convergence of the association rate constants k and the transition event duration distributions $F(t_{ed})$ obtained from the weighted ensemble simulations. However, self-convergence of these metrics does not guarantee that the simulation has converged to the true value of k or $F(t_{ed})$.

As an illustration, consider the convergence of $F(t_{ed})$, the probability distribution of the event duration times Ft_{ed} . Even if two transition event distributions $F_{\tau(1)}(t_{ed})$ and $F_{\tau(2)}(t_{ed})$ obtained by time points $N_{\tau(2)}$ and $N_{\tau(1)} > N_{\tau(1)}$ in a weighted ensemble simulation are statistically equivalent, this does not necessarily indicate asymptotic convergence on the true transition event duration distribution. Because a weighted ensemble simulation of length iterations contains only trajectories of maximum length τN_{τ} , then the statistical equivalence of $F_{\tau(1)}(t_{ed})$ and $F_{\tau(2)}(t_{ed})$ does not indicate that the entire event duration distribution has been adequately sampled, merely that all pathways taking time $t \leq \tau N_{\tau(1)}$ to traverse have been adequately sampled. Thus, for a weighted ensemble simulation of length

τN_τ , one must ultimately decide whether data obtained for time scales less than τN_τ are sufficient to provide insights into the systems under study.

4.4. How Does One Choose Optimal Weighted Ensemble Parameters? Efficient use of weighted ensemble sampling involves finding the optimal balance between computational expense and level of sampling. A poor choice of progress coordinate bins can easily lead to oversampling relatively unimportant regions of phase space. A large number of replicas not only aids rapid exploration of phase space but also determines the precision of probability current value and thus kinetic information; however, the total computational cost of weighted ensemble scales approximately linearly with the maximum number of system replicas. A short propagation/resampling period τ allows many opportunities for replicas to split and explore newly visited regions of phase space and for replicas to merge to avoid oversampling regions of phase space but ultimately may not allow sufficient divergence of trajectories to allow for efficient exploration of phase space.

Integral to the construction of a weighted ensemble simulation is the choice of a progress coordinate that is sufficiently sensitive to quantify “how far along” the reaction is. Any number of relatively low-cost enhanced-sampling or energy landscape smoothing techniques^{57–59} might be employed to guide the choice of a progress coordinate, including metadynamics;^{60,61} targeted,⁶² steered,⁶³ or accelerated⁶⁴ molecular dynamics; or the recently developed orthogonal space random walk method.⁶⁵ A number of short brute force simulations may be required to determine the average time evolution of the progress coordinate, which in turn determines the most efficient choices of bin spacing and the propagation/reweighting period τ . Finally, it may be necessary to adjust these parameters “on the fly” during a simulation, especially for large systems with complex, rough energy landscapes (i.e., proteins) where long-lived intermediate states may be encountered in the course of a simulation.

The complexities and advantages of actively adjusting the numbers of bins, their boundaries, and the number of replicas in each bin have been discussed in detail;² such schemes could be used to detect replicas that “stall” in certain progress coordinate bins and adjust the weighted ensemble simulation to compensate. These schemes would not be able to cope effectively with systems possessing intermediate states with lifetimes comparable to the mean first passage time; such systems do not exhibit the separation of time scales which weighted ensemble sampling is designed to exploit. However, using ideas developed from nonequilibrium umbrella sampling, it is possible to reweight phase space density analytically in order to accelerate the attainment of steady-state probability recycling;¹⁷ this would in turn accelerate the determination of the rate constant in systems with $t_{\text{ed}} \approx t_{\text{fp}}$ at the possible expense of efficient sampling of the transition path ensemble.

Finally, it should be noted that the weighted ensemble approach is but one instance of a class of “interface-based” enhanced sampling techniques which share a number of strengths and potential weaknesses;^{11,66,67} other such techniques include transition interface sampling (TIS) and variants,^{5,6,10} forward flux sampling (FFS),^{8,9} and milestoning.⁷ All of the methods in this class are rare event sampling methods that divide phase space along distinct interfaces, and each method is capable of providing realistic kinetic rates. Provided a well-chosen progress coordinate, these methods are equivalent in principle with respect to the information which can be obtained from them and

the efficiency with which that information is obtained, at least for equilibrium systems. Among these methods, however, the weighted ensemble approach is uniquely flexible; in particular, sampling can be maximized while minimizing computational cost both by dividing phase space according to arbitrary boundaries in any number of dimensions and by adjusting the level of sampling within each region (by adjusting the number of simulation replicas within a bin). The cost of this flexibility, however, is the complexity of determining efficient choices for parameters such as the progress coordinate, bin boundaries, and the number of replicas per bin. In situations where a reasonable progress coordinate cannot be determined, a method not dependent on a progress coordinate (such as transition path sampling^{3,68,69} or a recently developed variation of milestoning⁷⁰) may be necessary. Similarly, if efficient choices for simulation parameters (such as bin boundaries and the number of replicas per bin) cannot be made in advance and adjustment of these parameters during a simulation is impractical, then a method like FFS (for which analytical expressions for efficiency as a function of simulation parameters exist^{71,72}) may be a better choice.

4.5. Why Are Efficiencies What They Are? The efficiency of a weighted ensemble simulation is largely determined by weighted ensemble simulation parameters, particularly the propagation/resampling period τ , the choice of progress coordinate(s), and the locations of bin boundaries.⁴² For some systems, brute force simulation is already highly efficient at sampling the molecular association events; this is confirmed by the modestly increased weighted ensemble sampling efficiencies (S_k and S_{ed}) for methane/methane, Na^+/Cl^- , and methane/benzene. However, the fact that the weighted ensemble approach increases rather than decreases efficiency indicates that, even in such cases, the weighted ensemble technique is capable of accelerating sampling of both k and $F(t_{\text{ed}})$. On the other hand, the very high relative efficiency of sampling in $\text{K}^+/\text{18-crown-6}$ ether is particularly encouraging. Despite the small size of the system, brute force MD was incapable of effective sampling of rate constants and transition event duration distributions for $\text{K}^+/\text{18-crown-6}$ ether, almost certainly due to the high (approximately $14 k_{\text{B}}T$, 8.3 kcal/mol) barrier to dissociation. Weighted ensemble sampling was able to obtain self-converged values of both the rate constant k and the transition event duration distribution $F(t_{\text{ed}})$. This is primarily because probability recycling completely circumvents the necessity to climb the $14 k_{\text{B}}T$ dissociation barrier in order to observe another binding event.

These results point encouragingly to the ability to simulate protein–protein binding events with weighted ensemble molecular dynamics. With well-chosen bin boundaries, the weighted ensemble technique should increase sampling efficiency exponentially with increasing barrier heights. This is because placing bin boundaries sufficiently close to each other effectively linearizes the probability of crossing a number of bins in succession, rather than surmounting a barrier in one step with a probability which decreases exponentially with barrier height.⁷³ As a concrete example, the barrier to association in a diffusion-limited protein–protein system is approximately $10 k_{\text{B}}T$ (roughly five times that of the model systems). If this exponential efficiency scaling holds, then one can expect about 20 000-fold improvement in sampling for such a system. In other words, if a given computational resource is otherwise capable of generating 500 ps per calendar day (a substantial but accessible level of computational power), this efficiency gain corresponds to reaching a time scale of about 1 ms in 100 days, compared to the 50 ns that would

otherwise be possible in the same amount of time. However, since protein–protein binding pathways involve significant metastable intermediate states (e.g., encounter complexes⁷⁴), it is possible for a simulation to “stall” in such a state. As discussed above, several techniques exist which may partially ameliorate this difficulty, but in the end, a number of simulations connecting the intermediate states may be necessary to fully explore binding events in such systems.

5. CONCLUSIONS

We have applied the weighted ensemble path sampling approach to molecular dynamics simulations in explicit solvent, enabling the detailed sampling of rare molecular association events. We have compared the efficiency of weighted ensemble sampling relative to brute force sampling in simulating association events of methane/methane, Na^+/Cl^- , methane/benzene, and K^+ /18-crown-6 ether. Relative to brute force simulation, weighted ensemble sampling of these four systems confirms that the weighted ensemble approach reproduces or even improves sampling of both the rate constant k and the distribution of transition event durations. This improvement is on the order of 300- and 1100-fold, respectively, for a system exhibiting significant conformational flexibility (K^+ binding with 18-crown-6 ether). We expect efficiency gains to grow with increasing barriers to association. However, the existence of significant metastable intermediate states may hinder sampling in such systems, requiring the use of various enhancements to the weighted ensemble method in order to explore binding events in such systems. Nonetheless, these results indicate that weighted ensemble sampling in conjunction with MD simulations is likely to allow for the effective determination of transition paths and rate constants for protein binding events.

■ ASSOCIATED CONTENT

S Supporting Information. OPLS/AA atom type assignments for methane/benzene and K^+ /18-crown-6 ether systems, brute force first passage time distributions and potential of mean force curves for model systems, discussion and movies of binding pathways for the K^+ /18-crown-6 ether system, bin boundaries for weighted ensemble simulations, and derivations of the efficiency metrics S_k and S_{ed} . This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: ltchong@pitt.edu.

■ ACKNOWLEDGMENT

We thank Dan Zuckerman and Divesh Bhatt (U. Pitt. Dept. of Computational Biology), Bin Zhang (U. Michigan Dept. of Chemistry), Michael Grabe and Josh Adelman (U. Pitt. Dept. of Biological Sciences), Gary Huber (UCSD Dept. of Bioengineering), and Karen Zwier and Jonathan Livengood (U. Pitt. Dept. of History and Philosophy of Science) for helpful discussion; we also thank Xianghong Qi for initial efforts. This work was supported by NSF CAREER award MCB-0845216 to L.T.C., a University of Pittsburgh Arts & Sciences Fellowship to M.C.Z., and a University of Pittsburgh Brackenridge Fellowship (underwritten by the United States Steel Foundation) to J.W.K.

■ REFERENCES

- (1) Henzler-Wildman, K. A.; Kern, D. *Nature* **2007**, *450*, 964.
- (2) Huber, G. A.; Kim, S. *Biophys. J.* **1996**, *70*, 97.
- (3) Dellago, C.; Bolhuis, P. G.; Csajka, F.; Chandler, D. *J. Chem. Phys.* **1998**, *108*, 1964.
- (4) Zuckerman, D. M.; Woolf, T. B. *Phys. Rev. E* **2000**, *63*, 1.
- (5) van Erp, T. S.; Moroni, D.; Bolhuis, P. G. *J. Chem. Phys.* **2003**, *118*, 7762.
- (6) Moroni, D.; Bolhuis, P. G.; van Erp, T. S. *J. Chem. Phys.* **2004**, *120*, 4055.
- (7) Faradjian, A. K.; Elber, R. *J. Chem. Phys.* **2004**, *120*, 10880.
- (8) Allen, R. J.; Warren, P.; ten Wolde, P. R. *Phys. Rev. Lett.* **2005**, *94*, 018104.
- (9) Allen, R. J.; Frenkel, D.; ten Wolde, P. R. *J. Chem. Phys.* **2006**, *124*, 024102.
- (10) van Erp, T. S. *Phys. Rev. Lett.* **2007**, *98*, 1.
- (11) Zwier, M. C.; Chong, L. T. *Curr. Opin. Pharm.* **2010**, *10*, 745.
- (12) Zhang, B. W.; Jasnow, D.; Zuckerman, D. M. *J. Chem. Phys.* **2010**, *132*, 054107.
- (13) Rojnuckarin, A.; Livesay, D. R.; Subramaniam, S. *Biophys. J.* **2000**, *79*, 686.
- (14) Rojnuckarin, A.; Kim, S.; Subramaniam, S. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 4288.
- (15) Zhang, B. W.; Jasnow, D.; Zuckerman, D. M. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 18043.
- (16) Bhatt, D.; Zuckerman, D. M. *J. Chem. Theory Comput.* **2010**, *6*, 3527.
- (17) Bhatt, D.; Zhang, B. W.; Zuckerman, D. M. *J. Chem. Phys.* **2010**, *133*, 014110.
- (18) Dang, L. X. *J. Chem. Phys.* **1994**, *100*, 9032.
- (19) Meng, E. C.; Kollman, P. A. *J. Phys. Chem.* **1996**, *100*, 11460.
- (20) Oostenbrink, C.; van Gunsteren, W. F. *Phys. Chem. Chem. Phys.* **2005**, *7*, 53.
- (21) Trzesniak, D.; van Gunsteren, W. F. *Chem. Phys.* **2006**, *330*, 410.
- (22) Thomas, A. S.; Elcock, A. H. *J. Am. Chem. Soc.* **2007**, *129*, 14887.
- (23) Trzesniak, D.; Kunz, A.-P. E.; van Gunsteren, W. F. *Chemphyschem* **2007**, *8*, 162.
- (24) Belch, A. C.; Berkowitz, M. L.; McCammon, J. A. *J. Am. Chem. Soc.* **1986**, *108*, 1755.
- (25) Dang, L. X.; Rice, J. E.; Kollman, P. A. *J. Chem. Phys.* **1990**, *93*, 7528.
- (26) Guàrdia, E.; Rey, R.; Padró, J. A. *Chem. Phys.* **1991**, *155*, 187.
- (27) Hummer, G.; Soumpasis, D.; Neumann, M. *Mol. Phys.* **1992**, *77*, 769.
- (28) Pratt, L. R.; Hummer, G.; Garcia, A. E. *Biophys. Chem.* **1994**, *51*, 147.
- (29) Koneshan, S.; Rasaiah, J. C. *J. Chem. Phys.* **2000**, *113*, 8125.
- (30) Patra, M.; Karttunen, M. *J. Comput. Chem.* **2004**, *25*, 678.
- (31) Baumketner, A. *J. Chem. Phys.* **2009**, *130*, 104106.
- (32) Fennell, C. J.; Bizjak, A.; Vlachy, V.; Dill, K. A. *J. Phys. Chem. B* **2009**, *113*, 6782.
- (33) Timko, J.; Bucher, D.; Kuyucak, S. *J. Chem. Phys.* **2010**, *132*, 114510.
- (34) Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M.; Tanabe, K. *J. Am. Chem. Soc.* **2000**, *122*, 3746.
- (35) Ringer, A. L.; Figs, M. S.; Sinnokrot, M. O.; Sherrill, C. D. *J. Phys. Chem. A* **2006**, *110*, 10822.
- (36) Dang, L. X.; Kollman, P. A. *J. Am. Chem. Soc.* **1990**, *112*, 5716.
- (37) Troxler, L.; Wipff, G. *J. Am. Chem. Soc.* **1994**, *116*, 1468.
- (38) Humphrey, W. *J. Mol. Graphics* **1996**, *14*, 33.
- (39) Hänggi, P.; Talkner, P.; Borkovec, M. *Rev. Mod. Phys.* **1990**, *62*, 251.
- (40) Efron, B. Y. B.; Tibshirani, R. *Stat. Sci.* **1986**, *1*, 54.
- (41) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in C*, 2nd ed.; Cambridge University Press: Cambridge, England, 1992.

- (42) Zhang, B. W.; Jasnow, D.; Zuckerman, D. M. *J. Chem. Phys.* **2007**, *126*, 074504.
- (43) Kolmogoroff, A. *Ann. Math. Stat.* **1941**, *12*, 461.
- (44) Kvam, P. H.; Vidakovic, B. *Nonparametric Statistics with Applications to Science and Engineering*; John Wiley & Sons: Hoboken, NJ, 2007.
- (45) Cambillau, C.; Bram, G.; Corset, J.; Riche, C.; Pascard-Billy, C. *Tetrahedron* **1978**, *34*, 2675.
- (46) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435.
- (47) Adelman, S.; Doll, J. *J. Chem. Phys.* **1976**, *64*, 2375.
- (48) Shirts, M. R.; Pitera, J. W.; Swope, W. C.; Pande, V. S. *J. Chem. Phys.* **2003**, *119*, 5740.
- (49) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577.
- (50) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J. Comput. Chem.* **1997**, *18*, 1463.
- (51) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684.
- (52) Schuler, L.; Daura, X.; van Gunsteren, W. F. *J. Comput. Chem.* **2001**, *22*, 1205.
- (53) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269.
- (54) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474.
- (55) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.
- (56) Chandler, D. *Introduction to Modern Statistical Mechanics*; Oxford University Press: New York, NY, 1987.
- (57) Elber, R. *Curr. Opin. Struct. Biol.* **2005**, *15*, 151.
- (58) Adcock, S. A.; McCammon, J. A. *Chem. Rev.* **2006**, *106*, 1589.
- (59) Lei, H.; Duan, Y. *Curr. Opin. Struct. Biol.* **2007**, *17*, 187.
- (60) Huber, T.; Torda, a. E.; van Gunsteren, W. F. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 695.
- (61) Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562.
- (62) Schlitter, J.; Engels, M.; Krüger, P. *J. Mol. Graphics* **1994**, *12*, 84.
- (63) Izrailev, S.; Stepaniants, S.; Balsera, M.; Oono, Y.; Schulten, K. *Biophys. J.* **1997**, *72*, 1568.
- (64) Hamelberg, D.; Mongan, J.; McCammon, J. A. *J. Chem. Phys.* **2004**, *120*, 11919.
- (65) Zheng, L.; Chen, M.; Yang, W. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 20227.
- (66) Allen, R. J.; Valeriani, C.; Rein ten Wolde, P. *J. Phys.: Condens. Matter* **2009**, *21*, 463102.
- (67) Escobedo, F. A.; Borrero, E. E.; Araque, J. C. *J. Phys.: Condens. Matter* **2009**, *21*, 333101.
- (68) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291.
- (69) Grünwald, M.; Dellago, C.; Geissler, P. L. *J. Chem. Phys.* **2008**, *129*, 194101.
- (70) Májek, P.; Elber, R. *J. Chem. Theory Comput.* **2010**, *6*, 1805.
- (71) Allen, R. J.; Frenkel, D.; ten Wolde, P. R. *J. Chem. Phys.* **2006**, *124*, 194111.
- (72) Borrero, E. E.; Escobedo, F. A. *J. Chem. Phys.* **2008**, *129*, 024115.
- (73) West, A. M. A.; Elber, R.; Shalloway, D. *J. Chem. Phys.* **2007**, *126*, 145104.
- (74) Gabdoulline, R. R.; Wade, R. C. *Curr. Opin. Struct. Biol.* **2002**, *12*, 204.

RNA Conformational Sampling: II. Arbitrary Length Multinucleotide Loop Closure

C. H. Mak,^{*,†} Wen-Yeuan Chung,[‡] and Nikolay D. Markovskiy[†]

[†]Department of Chemistry, University of Southern California, Los Angeles, California 90089-0482, United States

[‡]Department of Mechanical Engineering, Chinese Cultural University, Taipei, Taiwan, Republic of China

ABSTRACT: In this paper, we describe how the inverse kinematic solution to the loop closure problem may be generalized to reclose a RNA segment of arbitrary length containing any number of nucleotides without disturbing the atomic positions of the rest of the molecule. This generalization is made possible by representing the boundary conditions of the closure in terms of a set of virtual coordinates called RETO, allowing the inverse kinematics to be reduced from the original six-variable/six-constraint problem to a four-variable/four-constraint problem. Based on this generalized closure solution, a new Monte Carlo algorithm has been formulated and implemented in a fully atomistic RNA simulation capable of moving loops of arbitrary lengths using torsion angle updates exclusively. Combined with other conventional Monte Carlo moves, this new algorithm is able to sample large-scale RNA chain conformations much more efficiently. The utility of this new class of Monte Carlo moves in generating large-loop conformational rearrangements is demonstrated in the simulated unfolding of the full-length hammerhead ribozyme with a bound substrate.

1. INTRODUCTION

The loop closure problem was first considered by Go and Scheraga in 1970 for polypeptide chains.¹ In biopolymers like proteins and nucleic acids, their bond lengths and bond angles are largely fixed. The flexibility of these molecules is therefore primarily derived from bond torsions. But for very long chains, even minute motions in a single torsion angle affect the coordinates of many atoms simultaneously, which may lead to excessive steric overlaps among the atoms being moved. For torsion angle moves in the conformational sampling of linear polymers to be practical, multiple torsion angles must be moved at the same time so that the atomic positions of the majority of the chain can remain relatively unperturbed. The loop closure problem seeks solutions for the possible sets of torsion angles inside a chain segment that have to be moved simultaneously in order to maintain the atomic coordinates of the rest of the chain fixed. Various formulations and extensions of the loop closure problem have been reported since Go and Scheraga's work.^{2–6} When applied to simulations of polymer conformational sampling, these loop closure solutions may be incorporated into a class of Monte Carlo (MC) schemes called concerted rotation or "conrot" moves.^{7–15} These concerted rotation algorithms have been successfully used in conformational sampling to generate trial moves for short segments along the backbone of linear polymers, proteins, and nucleic acids. A related set of methods called "rebridging MC"^{16,17} have also been used previously to reclose peptide chains as well as small ring structures.

In the original formulation of Go and Scheraga, they considered the case of a protein where the rotatable (ϕ, ψ) angles are separated by fixed peptide bonds in the trans conformation. They concluded that at least six torsion angles belonging to a tripeptide sequence must be moved simultaneously in order for the atomic coordinates of the rest of the chain to remain fixed. Dodd et al.⁴ and Deem and Bader⁹ used this to develop enhanced MC schemes to update the

conformation of short segments in linear polymer chains and proteins. Later, Dinner¹⁰ extended this formulation to allow arbitrary values for the fixed intervening torsion angles between the rotatable ones. Extensions of this formulation have also been applied to nucleic acids.^{10,15} Different formulations of the same loop closure problems were also reported recently for proteins by Coutsias et al.^{6,18} and for single nucleotides and ribose by Mak.¹⁹

In this paper, we consider the closure of an arbitrarily long segment on the backbone of a linear polymer. We generalize the closure problem so that it may be applied to reclose loops of any length. This generalized formulation applies to structures from as small as a single nucleotide in a RNA to loops of any length with fixed intervening torsion angles within three connecting segments. This generalized formulation relies on the reduction of the original six-variable/six-constraint problem to a four-variable/four-constraint problem, and a simple solution is derived using geometrics of rigid bodies well-known in the field of the kinematics of mechanisms. In various limits, our formulation is related to the conrot algorithm^{7–15} as well as the rebridging MC method.^{16,17} The generalized formulation allows for an efficient numerical or analytical solution. We show how this generalized closure solution may be used to devise several MC moves that are able to sample large-scale loop conformational changes in RNAs. We demonstrate the utility of these new MC moves in an all-atom simulation, studying the possible unfolding pathways of a ribozyme containing 63 nucleotides.

2. SINGLE-NUCLEOTIDE CLOSURE AND THE RETO COORDINATES

We begin with a brief review of the single-nucleotide RNA closure problem. This has been considered by one of us in a

Received: December 2, 2010

Published: March 11, 2011

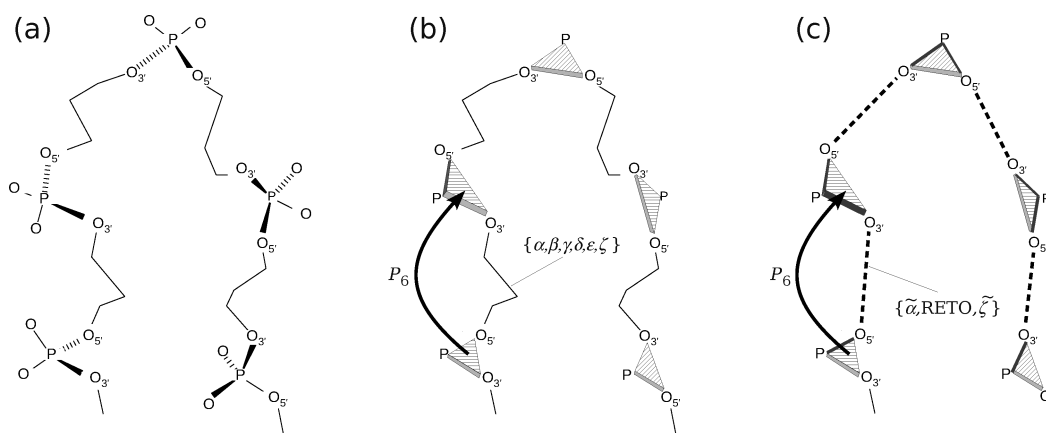


Figure 1. (a) Drawing shows a portion of the nucleic acid backbone. Starting from the 5' end, given the coordinates of the first phosphate group, the coordinates of all subsequent phosphate groups are uniquely determined if all intervening torsion angles are known. (b) Since bond lengths and bond angles are fixed, the coordinates of all five atoms in each phosphate group are uniquely determined if only the position of P and the orientation of the triangle O_3PO_5 are known. From one phosphate triangle, the position and orientation of the next phosphate triangle is known by either specifying the six intervening torsion angles, $\{\alpha, \beta, \gamma, \delta, \epsilon, \zeta\}$, or by P_6 . (c) Starting from one phosphate triangle, the position and orientation of the next phosphate triangle is uniquely determined if the six variables $\{\tilde{\alpha}, \text{RETO}, \tilde{\zeta}\}$ are specified. These are represented by dashed lines.

previous paper,¹⁹ which we shall refer to as paper (I). Figure 1a shows several nucleotides along the backbone of a nucleic acid. For clarity, the side chains and sugar rings have been removed for the drawing. In the simplest variant of the loop closure problem, we assume all bond lengths and bond angles are fixed and consider bond torsions as the only motions in the molecule. (This assumption may easily be relaxed if desired.) Starting from the 5' end, if the position and orientation of the first phosphate group is known, then the positions and orientations of all subsequent phosphate groups can be uniquely determined if all intervening torsion angles are specified. As such, the coordinates of every atom along the backbone of the chain may be easily reconstructed based on a knowledge of the coordinates of the first phosphate group and all the backbone torsion angles. In the case of RNAs, the forward kinematic problem is to solve for the positions of all the phosphate groups given the backbone torsion angles. Clearly, the forward kinematic problem of transforming from torsion angles to real-space coordinates is rather trivial.

The inverse kinematic problem of transforming from real-space coordinates back to torsion angles is much more involved. We can phrase the problem this way: If the coordinates of each phosphate group along the chain are specified, is it possible to solve for the intervening torsion angles given that the bond lengths and bond angles are rigid?

To begin, we recognize first that while every phosphate group has five atoms, the coordinates of all of them are uniquely determined if only the position of P and the orientation of the O_3PO_5 triangle is known. For convenience, we will call this the “phosphate triangle”. The coordinates of the rest of the atoms in this phosphate group are then fixed by the rigid bond lengths and bond angles. The phosphate triangles are shown in Figure 1b, where each phosphate group has been replaced by an oriented phosphate triangle. Using this reduction, the inverse kinematic problem may be restated alternatively as: If the position and orientation of every phosphate triangle is specified, is it possible to solve for all the torsion angles along the chain?

To be a mathematically well-posed problem, the number of degrees of freedom must be greater than or equal to the number of constraints. Six constraints (three translations and three Euler

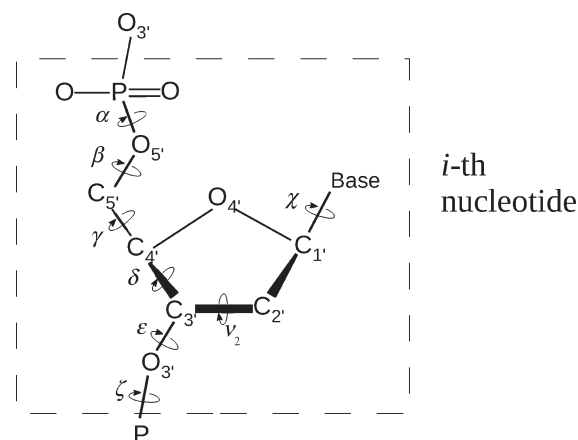


Figure 2. The six torsion angles along the backbone of a nucleotide are conventionally labeled α to ζ . For simplicity, O_2' has been omitted from the drawing.

angles) are involved in specifying the position and orientation of each phosphate triangle relatively to the previous one. We will call these six variables P_6 . For the inverse kinematic problem to be solvable, there must be a minimum of six torsion angles in each nucleotide. Interestingly, there are exactly six backbone torsion angles inside each nucleotide. These six torsion angles, shown in Figure 2, are conventionally labeled α to ζ . Consequently, the smallest reclosable loop in a RNA is a single nucleotide, and the inverse kinematic problem is to seek the transformation $P_6 \rightarrow \{\alpha, \beta, \gamma, \delta, \epsilon, \zeta\}$. This is known as the 6R problem in robotics.²⁰ (If there were fewer than six free torsion angles in a nucleotide, the inverse kinematic problem can no longer be phrased in terms of the phosphate triangles, but the problem can still be solved for any segments with six torsion angles.)

To further simplify the loop closure problem, paper (I) shows that the position and orientation of one phosphate triangle relative to the previous one may be given in terms of a new set of internal coordinates. These internal coordinates are defined in Figure 3, where r is the distance from O_5' to the next O_3' , η is the

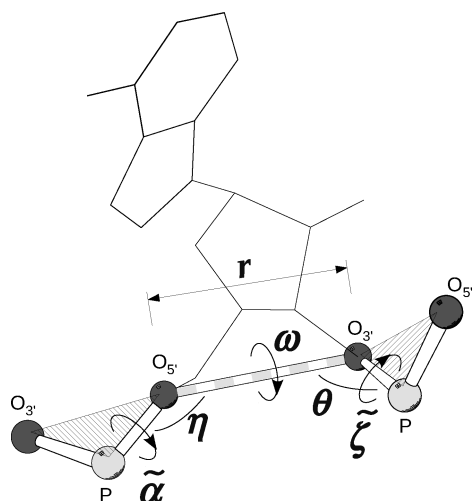


Figure 3. Internal coordinates used in the RETO representation to describe the relative position and orientation of two consecutive phosphate triangles along the backbone of a RNA. (Portions of this and other figures in this paper were generated using molscript).²¹

$P-O_5'-O_3'$ virtual bond angle, θ is the $O_5'-O_3'-P$ virtual bond angle, and ω is the $P-O_5'-O_3'-P$ virtual torsion angle. These four variables, $\{r, \eta, \theta, \omega\}$, are given the special name “RETO coordinate” because they play a special role in the generalized loop closure solution. The four atoms, P, O_5' , O_3' and the next P, are said to form a structure called a “RETO element”. The RETO coordinates are simply the minimal set of variables needed to uniquely specify the internal geometry of any RETO element. We also define $\tilde{\alpha}$ to be the $O_3'-P-O_5'-O_3'$ virtual torsion angle on the 5' end outside the RETO element, and $\tilde{\zeta}$ to be the $O_5'-O_3'-P-O_5'$ virtual torsion angle on the 3' end. Together, this set of six variables, $\{\tilde{\alpha}, \text{RETO}, \tilde{\zeta}\}$, specifies the position and orientation of one phosphate triangle relative to the previous one. This is shown in Figure 1c, where the phosphate groups are depicted as connected RETO elements, and their relative positions and orientations are now specified by the RETO variables and the torsion angles $\tilde{\alpha}$ and $\tilde{\zeta}$. This is shown in greater detail in Figure 3, which illustrates that starting from one phosphate triangle, specifying $\{\tilde{\alpha}, \text{RETO}, \tilde{\zeta}\}$ will uniquely determine the position and orientation of the next phosphate triangle. In terms of the new RETO coordinates, the inverse kinematic problem is now to seek the transformation $\{\tilde{\alpha}, \text{RETO}, \tilde{\zeta}\} \rightarrow \{\alpha, \beta, \gamma, \delta, \varepsilon, \zeta\}$, which corresponds to the mapping from Figure 1c back to 1b.

Paper (I) shows that with the introduction of the RETO coordinates, the original six-constraint/six-variable problem may be reduced to a simpler four-constraint/four-variable problem. This is because the variable $\tilde{\alpha}$ maps directly onto α , while $\tilde{\zeta}$ maps onto ζ . In mathematical terms, the Jacobian matrix (J) of the transformation $\{\tilde{\alpha}, \text{RETO}, \tilde{\zeta}\} \rightarrow \{\alpha, \beta, \gamma, \delta, \varepsilon, \zeta\}$:

$$J = \frac{\partial(\tilde{\alpha}, \text{RETO}, \tilde{\zeta})}{\partial(\alpha, \beta, \gamma, \delta, \varepsilon, \zeta)} \quad (1)$$

turns out to be block diagonal, with $\tilde{\alpha}$ forming a 1×1 block with α , RETO forming a 4×4 block with $\{\beta, \gamma, \delta, \varepsilon\}$, and $\tilde{\zeta}$ forming a 1×1 block with ζ . Furthermore, it is easy to show that the 1×1 block between $\tilde{\alpha}$ and α is unity for all values of $\tilde{\alpha}$, and the same is true for the 1×1 block between $\tilde{\zeta}$ and ζ .

A detailed solution for the inverse kinematic transformation $\text{RETO} \rightarrow \{\beta, \gamma, \delta, \varepsilon\}$ has been given in paper (I). Briefly, the

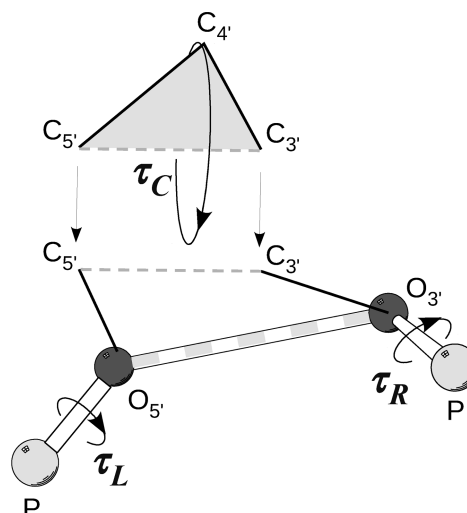


Figure 4. Drawing illustrating the single-nucleotide closure solution. RETO is fixed by the positions of P, O_5' , O_3' , and the next P. The chain is divided into three rigid elements: O_5' with its bonds to P and C_5' form the left element, C_4' with its bonds to C_5' and C_3' the center element, and O_3' and its bonds to C_3' and P the right element. The closure is obtained by the rotations τ_L and τ_R to match the native $C_5'-C_3'$ distance, the reattachment of the center element, and the rotation τ_C around the $C_5'-C_3'$ axis to produce the correct $O_5'-C_5'-C_4'$ bond angle.

solution is illustrated in Figure 4 and it proceeds as follows: Given the RETO coordinates, the relative positions of P, O_5' , O_3' , and the next P are fixed. With the known $P-O_5'$ and $O_5'-C_5'$ bond lengths and the $P-O_5'-C_5'$ bond angle, we can consider the three atoms P, O_5' , and C_5' as a rigid element (the left element). Similarly, with the known $C_3'-O_3'$ and $O_3'-P$ bond lengths and the $C_3'-O_3'-P$ bond angle, we can also consider the three atoms C_3' , O_3' , and P as another rigid element (the right element). Finally, the atom C_4' and its two adjacent bonds to C_5' and C_3' may be considered as yet another rigid element (the center element), which is shown in Figure 4 as the gray triangle. To construct the closure solution, we imagine first detaching the center element from the loop. The left element is then free to rotate about the $P-O_5'$ bond through some rotation angle τ_L . Similarly, the right element can also rotate about the $P-O_3'$ bond through angle τ_R . In order for the center element to be reattached properly, the distance between C_5' and C_3' , shown as a gray dashed line between the left and right elements, must match the $C_5'-C_3'$ distance on the lower edge of the gray triangle of the center element. If we consider τ_L as the input, then the output angle τ_R that produces the correct $C_5'-C_3'$ distance will form a discrete set of points for every input τ_L . For each one of these τ_R values, we can then reattach the center element. But in order to also produce the correct $O_5'-C_5'-C_4'$ bond angle, the center element must also be rotated about the $C_5'-C_3'$ axis by the proper angle(s) τ_C . After this, we can measure the $C_4'-C_3'-O_3'$ angle. If this matches the correct $C_4'-C_3'-O_3'$ angle, then the closure is solved. The solution of the closure problem is therefore obtained by expressing the output $C_4'-C_3'-O_3'$ angle as a function of the input τ_L . This is followed by a root search to determine the value(s) of τ_L which produce(s) the correct angle matching the native $C_4'-C_3'-O_3'$ angle. After the solution of this four-constraint/four-variable problem is obtained, the last two torsion angles $\tilde{\alpha}$ and $\tilde{\zeta}$ are easily determined since they are simply equal to their counterparts α and ζ , plus some offsets.

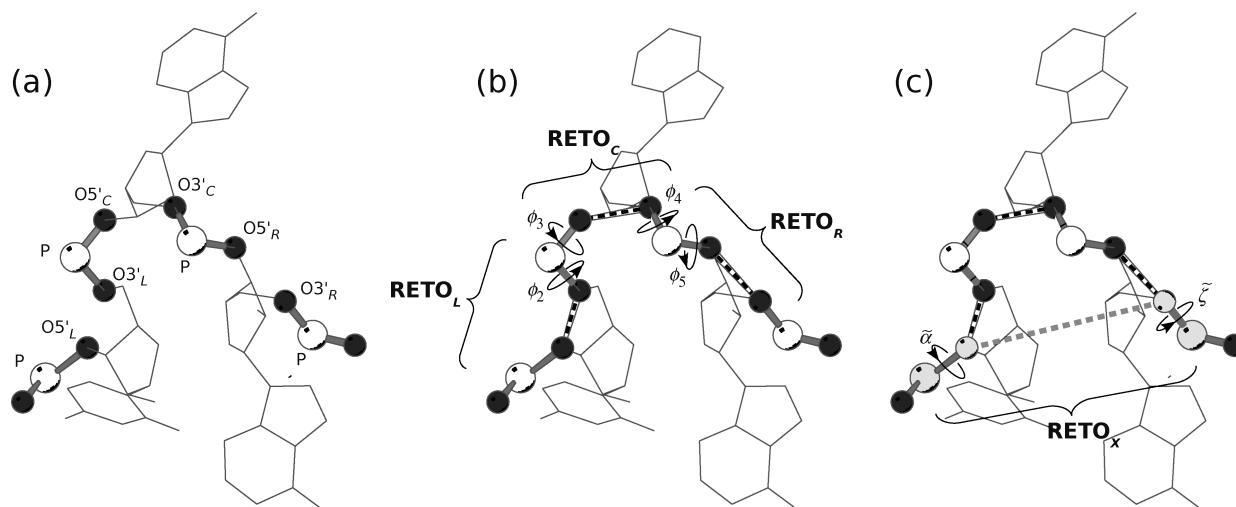


Figure 5. The three-nucleotide loop closure problem. (a) Drawing showing phosphate triangles in three consecutive nucleotides. (b) The RETO representation of the left, center and right elements, each one containing a one-nucleotide segment. Their relative orientations are defined by four intervening virtual torsion angles: ϕ_2 , ϕ_3 , ϕ_4 , and ϕ_5 . (c) The boundary conditions are specified by RETO_X , which defines the geometry of the four boundary atoms shown in gray for this three-nucleotide sequence. The orientations of the rest of chain to the left (the 5' end) and the right (the 3' end) are specified by two additional virtual torsion angles α and ζ .

In addition to the backbone torsion angles, the ribose also has internal torsions. These are coupled to the mainchain torsion angle δ . When the nucleotide backbone is reclosed, δ is modified, which requires the ribose to be reclosed simultaneously. This problem of reclosing the ribose is analogous to the conformation of prolines in a polypeptide, which has been considered by Ho et al.¹⁸ who shows that the standard loop closure solution may be applied with slight modification to reclose any five-membered ring structure. As such, one can simply use the same closure algorithm to reclose the ribose, and full details have been given in paper (I).

3. MULTINUCLEOTIDE CLOSURE

The inverse kinematic solution for reclosing a RNA backbone sequence containing more than a single nucleotide is a straightforward extension of the single-nucleotide closure solution. The extension is facilitated by the RETO coordinates described in the last section. To illustrate how this works, we will start with the closure of a three-nucleotide loop, a schematic of which is shown in Figure 5a. The P atoms are colored white, and the O atoms are colored black. For clarity, only the two phosphate oxygens on the backbone are shown. The rest of the chain is displayed as sticks. The 5' end of the chain is on the left. We identify the O atoms on the leftmost nucleotide by the subscript L, those on the center nucleotide by C, and the ones on the rightmost nucleotide by R.

Figure 5b shows how this three-nucleotide sequence may be divided into three connected RETO elements. RETO_L on the left consists of the P, $O_{5'}$, and $O_{3'}$ atoms from the first nucleotide plus the P atom from the second nucleotide. RETO_C in the center consists of P, $O_{5'}$, and $O_{3'}$ from the second nucleotide plus P from the third and similarly with RETO_R on the right. Two consecutive RETO elements are connected through the P atom they share. In Figure 5b, we have indicated the virtual bond between $O_{5'}$ and $O_{3'}$ in each RETO element using a thick dashed line. As described in Section 2, the geometry of each RETO element is specified by its four RETO variables $\{r, \eta, \theta, \omega\}$.

In addition to defining the RETO variables of the three internal elements, we can also define a set of RETO variables specifying the relative positions of the four terminal atoms, P and $O_{5'}$ on the 5' end and $O_{3'}$ and P on the 3' end. We called these the RETO_X variables. The four atoms that define RETO_X are colored gray in Figure 5c. Similar to the single-nucleotide case in Figure 3, two additional virtual torsion angles α and ζ are needed to determine how RETO_X is oriented relative to the rest of the chain outside.

If the internal structure of each RETO element is fixed, the complete conformation of this three-nucleotide sequence can be uniquely reconstructed if we specify the RETO coordinates of each element as well as the two virtual torsion angles between each pair of connected RETO elements. These torsion angles are depicted in Figure 5b. The two virtual torsion angles connecting RETO_L and RETO_C are ϕ_2 , the $O_{5'}-O_{3'}-P-O_{5'}$ angle, and ϕ_3 , the $O_{3'}-P-O_{5'}-O_{3'}$ angle. Similarly, the two virtual torsion angles connecting RETO_C and RETO_R are ϕ_4 , the $O_{5'}-O_{3'}-P-O_{5'}$ angle, and ϕ_5 , the $O_{3'}-P-O_{5'}-O_{3'}$ angle. Together with RETO_L , RETO_C , and RETO_R , specifying these four virtual torsion angles will uniquely define the complete conformation of the chain, producing the final positions of $O_{3'}$ and P on the 3' end relative to the positions of P and $O_{5'}$ on the 5' end.

In terms of the variables defined above, the inverse kinematic problem of closing a three-nucleotide loop is to seek the transformation $\text{RETO}_X \rightarrow \{\phi_2, \phi_3, \phi_4, \phi_5\}$, given known RETO_L , RETO_C , and RETO_R and fixed bond lengths and bond angles. Once this problem is solved, the two virtual torsion angles and ζ in Figure 5c are then trivially given by the native torsion angle α on the 5' end and ζ on the 3' end plus some offsets, just as in the single-nucleotide closure problem.

The solution of the problem $\text{RETO}_X \rightarrow \{\phi_2, \phi_3, \phi_4, \phi_5\}$ follows closely the single-nucleotide solution described in the last section. The only difference is that the left, center, and right rigid elements are now the RETO_L , RETO_C and RETO_R in Figure 5, instead of the three-atom segments shown in Figure 4. Other than this, the rest of the solution is identical to the single-nucleotide case, and its details will not be repeated again.

At this point, it should also be clear that the solution described above is applicable to not only a three-nucleotide loop, but it can also be applied to close loops of any size. To close larger loops, the only modification needed is to redefine the RETO_L , RETO_C , or RETO_R elements so that each may encompass more than one nucleotide. For the three-nucleotide closure above, each of the RETO_L , RETO_C , and RETO_R elements contain only one nucleotide, and we may call this a 1 + 1 + 1 closure. To close larger loops, such a nine-nucleotide loop for instance, we may take each of RETO_L , RETO_C , and RETO_R to be a rigid three-nucleotide sequence and perform a 3 + 3 + 3 closure. But the same nine-nucleotide loop may also be closed using a 2 + 5 + 2 or a 2 + 6 + 1 closure or any other combination of sizes. Furthermore, if desired, the closure may even be applied in a cascading or a recursive manner. For example, we may carry out a 1 + 1 + 1 closure for each of RETO_L , RETO_C , and RETO_R and then reclose the composite nine-nucleotide loop by a second 3 + 3 + 3 closure. Clearly, by using different combinations of single-nucleotide and $l_L + l_C + l_R$ multinucleotide closures, we may reclose loops of any arbitrary length.

4. MONTE CARLO ALGORITHMS USING MULTINUCLEOTIDE CLOSURE

We can incorporate the multinucleotide closure solution above to construct a set of novel MC moves that are capable of updating the conformation of an arbitrarily long RNA sequence. We will consider three variants of this MC method, which differ in complexity. In certain limits, they are related to the conrot algorithm^{7–15} in various ways, but a driver angle is not used in our method. We will describe each MC algorithm by referring to the drawings in Figure 5, but it is important to remember that any of the RETO_L , RETO_C , and RETO_R elements may contain more than one nucleotide.

4.1. Monte Carlo Variant MC1. This is the simplest of all three MC variants. In this MC move, we will keep the RETO_X geometry as well as all three internal RETO elements (L, C, and R) rigid and let the closure select a new solution to reclose the loop into a different conformation. Since we are not altering RETO_X , the rest of the chain to the left on the 5' end of the drawing in Figure 5 as well as to the right on the 3' end are fixed during this MC move. Because the internal structures of RETO_L , RETO_C , and RETO_R are also frozen, the reclosing of the loop may easily be done by randomly picking one closure solution from among the possible sets of $\{\phi_2, \phi_3, \phi_4, \phi_5\}$ that solve the closure problem with RETO_X as the boundary condition. Because the constraints are in the RETO_X instead of in torsion angle space, a Jacobian is involved. The Jacobian that is needed is

$$J = \left| \frac{\partial(\phi_2, \phi_3, \phi_4, \phi_5)}{\partial(r_X, \eta_X, \theta_X, \omega_X)} \right| \quad (2)$$

for the transformation $\{\phi_2, \phi_3, \phi_4, \phi_5\} \rightarrow \text{RETO}_X$. This is more easily computed from its reciprocal $|\partial(r_X, \eta_X, \theta_X, \omega_X)/\partial(\phi_2, \phi_3, \phi_4, \phi_5)|$ using numerical differentiation. This MC move therefore consists of three simple steps:

- (1) Randomly select a residue as the starting point of the RETO_L element. Assign the next l_L nucleotides to the RETO_L element, l_C nucleotides to RETO_C , and then l_R nucleotides to RETO_R .

- (2) With RETO_X , RETO_L , RETO_C , and RETO_R fixed, use a multinucleotide closure to generate all loop solutions. Randomly select one of the solutions to reclose the loop.
- (3) Accept or reject the new loop conformation based on its energy compared to the energy of the old state using the Metropolis²² rule:

$$P = \min \left[1, \frac{J' \exp(-E'/kT)}{J \exp(-E/kT)} \right] \quad (3)$$

where E is the energy of the old conformation, E' is the energy of the new, J and J' are the Jacobians of the new and old conformations, respectively, k is Boltzmann's constant, and T is the temperature.

Clearly, steps 1 and 2 generate the trial state with a symmetric transition probability. Together with the Metropolis acceptance criterion in step 3, this MC strictly satisfies detailed balance. The lengths of each of the three RETO elements, l_L , l_C , and l_R may either be fixed or chosen randomly.

4.2. Monte Carlo Variant MC2. In the second MC variant, we allow the internal RETO elements to be flexible while keeping RETO_X fixed. But instead of moving the variables $\{r, \eta, \theta, \omega\}$ in each of RETO_L , RETO_C , and RETO_R directly, we simply apply random displacements to each of their internal torsion angles to arrive at a new RETO geometry for each element. For instance, when applying this MC variant to a four-nucleotide sequence using a 1 + 1 + 2 closure, we would randomly displace the β , γ , δ , and ε torsion angles inside the RETO_L and RETO_C elements first. This will generate new RETO_L and RETO_C geometries. Then for the two-nucleotide-long RETO_R , we will displace all torsion angles between β of the first residue in RETO_R and ε in the last residue of RETO_R . This would result in new geometries for all three internal RETO elements, and we would then reclose the loop for the fixed RETO_X on the outside into a new loop conformation. Since the density of conformational states is uniform in torsion angle space, generating displacements for RETO_L , RETO_C , and RETO_R this way will not require an additional Jacobian other than the one already used in MC1. This MC move therefore consists of the following steps:

- (1) Randomly select a residue as the starting point of the RETO_L element. Assign the next l_L nucleotides to the RETO_L element, l_C nucleotides to RETO_C , and then l_R nucleotides to RETO_R .
- (2) Using the current geometries of RETO_L , RETO_C , and RETO_R , obtain the number of loop closure solutions N with boundary condition fixed by RETO_X .
- (3) Apply random displacements to all internal torsion angles in each of the RETO_L , RETO_C , and RETO_R elements.
- (4) With the new geometries for the RETO_L , RETO_C , and RETO_R elements, use a multinucleotide closure to generate all new loop solutions for the fixed RETO_X . Let the number of new solutions be N' .
- (5) Randomly select from one of the new solutions to reclose the loop.
- (6) Accept the new loop conformation using the probability:

$$P = \min \left[1, \frac{N'J' \exp(-E'/kT)}{NJ \exp(-E/kT)} \right] \quad (4)$$

It can be shown easily that this MC scheme strictly satisfies detailed balance. The lengths of each of the three RETO elements, l_L , l_C , and l_R may either be fixed or chosen randomly.

4.3. Monte Carlo Variant MC3. In this third and final MC variant, we keep the three internal RETO elements (L , C , and R) frozen and allow only RETO_X to move. To move RETO_X , we have no alternative but to displace $\{r_X, \eta_X, \theta_X, \omega_X\}$ directly. So we would reclose the loop using the original RETO_L , RETO_C , and RETO_R elements onto the new RETO_X constraint. This MC move therefore consists of the following steps:

- (1) Randomly select a residue as the starting point of the RETO_L element. Assign the next l_L nucleotides to the RETO_L element, l_C nucleotides to RETO_C , and then l_R nucleotides to RETO_R .
- (2) Using the current geometries of RETO_L , RETO_C , and RETO_R , obtain the number of loop closure solutions N with boundary condition fixed by RETO_X .
- (3) Generate a trial RETO_X by applying random displacements to $\{r_X, \eta_X, \theta_X, \omega_X\}$.
- (4) Displace the rest of the molecule on the 5' end external to RETO_X and on the 3' end so that their relative positions and orientations are consistent with the trial RETO_X . Then apply random displacements to \hat{r} , which connects RETO_X to the rest of the molecule on the 5' end, and to \hat{c} , which connects RETO_X to the rest of the molecule on the 3' end.
- (5) Use a multinucleotide closure to generate all new loop solutions for the new RETO_X using the old RETO_L , RETO_C , and RETO_R geometries. Let the number of new solutions be N' .
- (6) Randomly select from one of the new solutions to reclose the loop.
- (7) Accept the new loop conformation using the probability:

$$P = \min \left[1, \frac{N'J' \exp(-E'/kT)}{NJ \exp(-E/kT)} \right] \quad (5)$$

It can be shown easily that this MC scheme strictly satisfies detailed balance. The lengths of each of the three RETO elements, l_L , l_C and l_R may either be fixed or chosen randomly.

5. DISCUSSION

While it is straightforward to construct MC moves based on the single- or multinucleotide closure solutions given above, the practical usefulness of these MC moves in an atomistic simulation is not guaranteed. In fact, our experience with loop-closure MC simulations shows that the effectiveness of MC moves based on loop closure is often severely limited by steric problems. If we start from an already folded conformational state in a long-chain biomolecule, then steric collisions will certainly prevent most of the new loop solutions from being accepted. Even though loop closure has the potential to generate large-scale conformational changes, the reality is that any trial loop conformation in a sterically congested region of the molecule will produce so many new steric overlaps that almost all trial conformations are rejected most all of the time.

As a test, we have used MC variant MC1 on single-nucleotide loops for a number of RNA structures in the PDB database. We discovered that the acceptance rate of any single-nucleotide loop conformations other than the native one is almost always zero. This result should not be too surprising. Any new loop conformation other than the native one almost always cause too many steric collisions, and if the loop is to be reclosed, then it will

almost always close back into the native conformation. Therefore, MC1 would be a very ineffective choice for moving single-nucleotide loop or very short loops. For these, MC2 or MC3 would be a better choice. In our MC simulations, we would use MC3 exclusively for single-nucleotide closures. However, single-nucleotide moves will only produce small-scale local motions that are very similar to the thermal motions typically seen in a molecular dynamics (MD) simulation. If an atomistic MC simulation of RNAs is based on single-nucleotide MC3 moves alone, then it will not be a very effective algorithm for studying large-scale loop motions. Multinucleotide MC moves must be added.

Using loop-closure MC moves to close larger loops introduces other problems. When we tried using MC2 or MC3 to reclose loops with nine or more nucleotides starting from the native structure of several RNAs in the PDB database, we discovered that not only was excessive steric collisions a frequent problem but also the loss of favorable base-pairing and base-stacking interactions when the structures of the RETO elements are altered leading to large energy costs. Therefore, while the MC moves MC2 and MC3 allow the RNA molecules to acquire larger conformational changes than MC1, the energy cost associated with this increased flexibility is mostly unfavorable. In fact, the larger the amplitude of the random displacements in RETO_L , RETO_C , RETO_R , or RETO_X we applied, the lower the acceptance rate became. So for reclosing loops larger than single-nucleotide ones, we relied exclusively on MC1, which is the simplest variant to implement numerically.

Since the majority of the new loop conformations are sterically impossible when long loops are reclosed, loop-closure MC simulations are intrinsically highly inefficient. If it was not for the fact that these MC moves possess the unique capability of generating large-scale motions that molecular dynamics or regular MC simulations cannot see, they would have been useless in a practical sense. In order to increase their efficiency and turn loop-closure MC into a practical simulation tool, we have considered various ways to try to speed up the computation. Since the calculation of the total energy is one of the most expensive operations in the algorithm, we tried to accelerate the closure moves by detecting excessive steric overlaps early and screening out sterically impossible closure solutions before their full energies were calculated. To do this, we used a minimum steric radius of 1.5 Å for each atom. If any pair of atoms in the new loop solution come closer than twice the minimum steric radius, then the configuration was rejected. This simple change dramatically reduced the overhead required for calculating the full energy for conformations which were sterically too costly. With this we were able to screen a much larger number of new loop conformations with a smaller amount of computational effort, making loop-closure MC a reasonably practical simulation method for finding large-scale loop motions.

Furthermore, when large loops are reclosed, the choice of the size of the L , C , and R RETO elements is arbitrary. Therefore, we are free to try to find combinations of l_L , l_C , and l_R that would optimize the performance of the loop closure MC. If MC1 is used, the three elements RETO_L , RETO_C , and RETO_R are fixed; therefore, the internal coordinates of all the atoms inside each element are fixed relative to each other. Since the closure solution rearranges these three elements, keeping each one as a rigid structure, any base pairing or base stacking interactions internal to any element will be preserved, while the base pairing or base stacking interactions between two different RETO elements may

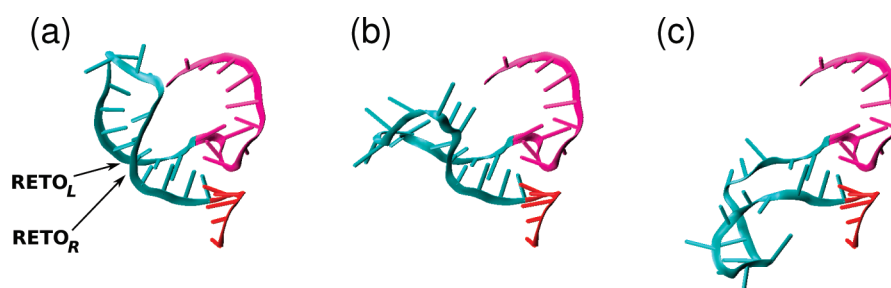


Figure 6. Sample conformations illustrating the effect of a $3 + R + 3$ MC1 move on a nucleotide segment: (a) The RETO_L and RETO_R elements randomly selected by this MC move. (b) New conformation generated by reclosing the loop with a rigid RETO_C in between. (c) Another new conformation generated using a similar loop-closure MC move.

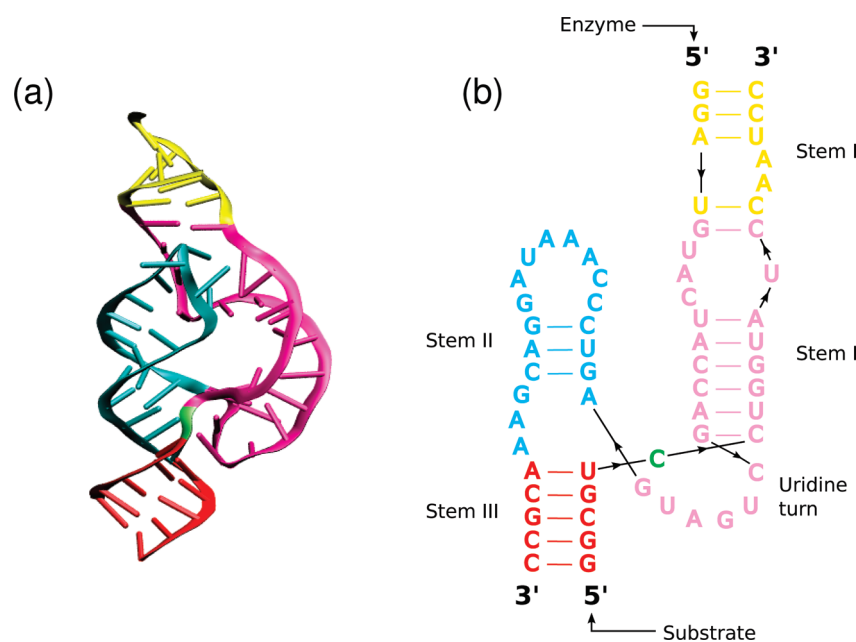


Figure 7. (a) Three-dimensional tertiary structure of the hammerhead ribozyme with a bound substrate (PDB code: 2GOZ). The substrate is the shorter chain, with its 5' end on the helix colored red (lower left of the figure), going through the nucleolytic site colored green, passing through the stem region colored purple, and finally ending in another stem region colored yellow (top of the figure). (b) Secondary structure corresponding to the three-dimensional structure in (a) using the color coding adopted from Martick and Scott.²³ (Portions of this and other figures in this paper were generated using VMD).²⁴

be disrupted when the loop is reclosed into a new conformation. Therefore, in order not to disturb the most favorable interactions, the L, C, and R RETO elements could be chosen so that there are no strong base pairing or base stacking interactions between different RETO elements. The details of how this idea may be implemented in a rigorous MC algorithm will be addressed in a forthcoming paper.

Since the main purpose of this paper is to demonstrate how loop closure solutions may be used to design a set of MC moves to study large-scale loop motions in RNAs, we will only present results from a MC simulation where l_L and l_R are preselected and fixed, while l_C is chosen randomly. In this simulation, we chose RETO_L and RETO_R to be three nucleotide long each. From our experience, the segment lengths $l_L = l_R = 3$ are long enough to provide the loop to be reclosed with adequate flexibility. In this way, the RETO_L and RETO_R elements are essentially acting as hinges that could allow the RETO_C element to unfold. Because closure requires at least one nucleotide on the left and right

outside of RETO_X and because RETO_L and RETO_R take up three nucleotide each, while RETO_C must have a minimum of one nucleotide, the length l_C can be chosen randomly according to $l_C = \text{int}(R \times [L - 9]) + 1$, where R is a uniformly distributed random number between 0 and 1, and L is the sequence length of the RNA. We will call this a $3 + R + 3$ MC1 move because it employs MC variant 1 with fixed lengths $l_L = l_R = 3$ for RETO_L and RETO_R and a random length l_C for RETO_C .

Figure 6 illustrates the unfolding of a stem loop in a small RNA fragment by a $3 + R + 3$ MC1 move. In every MC pass of a $3 + R + 3$ MC1 move, the first residue of RETO_L as well as the length l_C were selected randomly according to the criteria given above. Corresponding to these choices, RETO_C and RETO_R elements were then determined. A loop closure MC move was then carried out using MC variant 1 described in Section 4.1. Figure 6b gives an example of a new loop conformation derived from a $3 + R + 3$ loop closure MC1 move, showing that the three-nucleotide RETO_L and RETO_R elements provide the loop

with sufficient conformational flexibility for it to unfold. Figure 6c shows another unfolded conformation of the same loop derived from the same loop-closure MC move.

6. EXAMPLE: UNFOLDING OF THE HAMMERHEAD RIBOZYME

In this section, we present MC simulation results to demonstrate how loop closure MC moves may be used to study large-scale loop motions in RNAs. The molecule we have chosen for the test is the full-length *Schistosoma* hammerhead ribozyme with a bound substrate, whose structure was determined recently by Martick and Scott²³ (PDB code: 2GOZ). The three-dimensional folded structure of the molecule is shown in Figure 7a with its secondary structure given in Figure 7b, and the color coding has been adopted from Martick and Scott's paper. Together, the ribozyme and substrate contain 63 nucleotides in two chains, A and B, with approximately 2000 atoms in total. The substrate is the shorter chain B, which in Figure 7a is positioned in the front. In the folded structure, the bulge in stem I interacts with the loop in stem II via a number of tertiary contacts.

The results described below were obtained from a 298 K simulation using 3 + R + 3 MC1 moves plus single-nucleotide MC3 moves with an all-atom model for the hammerhead as well as the substrate. Potential energies were evaluated using the Amber ff98 force field^{25,26} and generalized Born/surface area (GB/SA)^{27,28} to model the solvent, with no explicit counterions. Without counterions, the native folded state may no longer be the most stable conformation in the simulation because counterions are known to be essential for RNA folding, and the native state may therefore unfold during the simulation.

To demonstrate that the loop closure MC we have proposed is useful for simulating large-scale loop motions in RNAs, we used it to investigate the possible unfolding motions of the hammerhead ribozyme. We carried out 256 independent simulations each starting with the same native structure. Each run consisted of 1000 MC passes, with one pass defined as the molecule having attempted one MC3 single-nucleotide move for each residue and $N/(3 + l_C + 3) 3 + l_C + 3$ MC1 moves for the entire molecule, where l_C was chosen randomly for each pass according to the procedure described in the last section, with the total sequence length of the molecule $L = 63$. At the end of each run, we examined the degree of unfolding of the final structure by computing its root-mean-square deviation (rmsd) from the native folded state using all atoms in the molecule.

The rmsd data from the simulations are summarized in Figure 8, which shows a histogram of all the results. While a majority of the final structures were very similar to the native folded state (with rmsd < 3 Å), there were also a significant number of final structures having rather different conformations from the native state. The final structures obtained from the MC simulations fall into several overlapping but identifiable clusters.

First, structures in cluster A with rmsd between 0 to 3 Å are those conformations that are similar to the native state. Structures in this cluster account for almost 70% of the final population. An example of a structure in this cluster is shown in Figure 9A.

The second cluster with the most distinctive structures is label E in Figure 8. These have rmsd between approximately 21 and 28 Å from the native state. A typical structure from this cluster is shown in Figure 9E, with the native conformation in translucent orange superposed on it. Structures in this cluster are

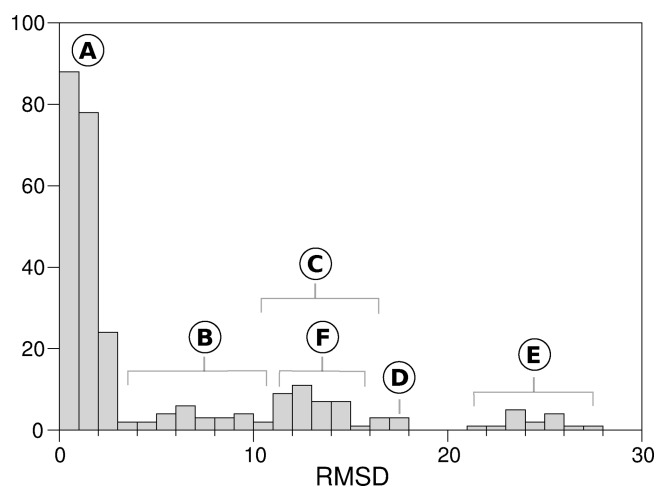


Figure 8. Histogram showing rmsd of final structures derived from 256 independent MC simulations of the hammerhead ribozyme starting from the native conformation. The simulations were carried out using only single-nucleotide MC3 moves plus 3 + R + 3 MC1 loop closure moves, with rigid bond lengths and bond angles. The final structures fall into six identifiable clusters, labeled A–F.

characterized by the complete unfolding of stem II in chain A, with one hinge ($RETO_L$) in the uridine turn (purple, chain A) and the other ($RETO_R$) in stem III (red, chain A). This class of structures may be important conformational intermediates which act as gateway states to the correct positioning of the enzyme for it to carry out the nucleolytic reaction it is to catalyze.

Structures in cluster B are similar to those in cluster E, but they are characterized by the partial instead of complete unfolding of the stem loop structure in stem II. An example is shown in Figure 9B, with the native structure in translucent orange superposed. For structures in this cluster, the unfolding of stem II is achieved with both hinges ($RETO_L$ and $RETO_R$) on stem II (cyan, chain A). The catalytic site in these structures are largely intact.

The rest of the clusters in Figure 8, namely C, D, and F, were derived from the native state via a very different type of loop motions. While the multinucleotide loop closure solution described in Section 4 has been developed to reclose a chain with a contiguous backbone, an interesting feature is that the same algorithm may actually be used to reclose a segment that spans two or more disjoint chains. For example, if the beginning and end of the $RETO_C$ element happen to be chosen such that $RETO_C$ straddles two different chains, reclosing the loop using MC1 will intrinsically preserve the relative coordinates of both segments in the interior of $RETO_C$, even though they belong to disjoint chains. In this way, if the molecule consists of two or more chains, rigid segments containing residues from multiple chains may be moved at the same time using MC1. Figure 9C and D shows two examples of structures derived from MC moves of this type. The structures belonging to clusters C and D in Figure 8 primarily come from the unfolding of stem I, which is formed from the hybridization of chains A and B. The example shown in Figure 9C came from the bending of a short section of stem I, using one hinge on chain A (purple) and another hinge on chain B (yellow). The one in Figure 9D involves the unfolding of stem I outside the uridine turn, with one hinge on chain A (purple) and the other on chain B (purple). Structures falling under cluster F in Figure 8 are also derived from a similar

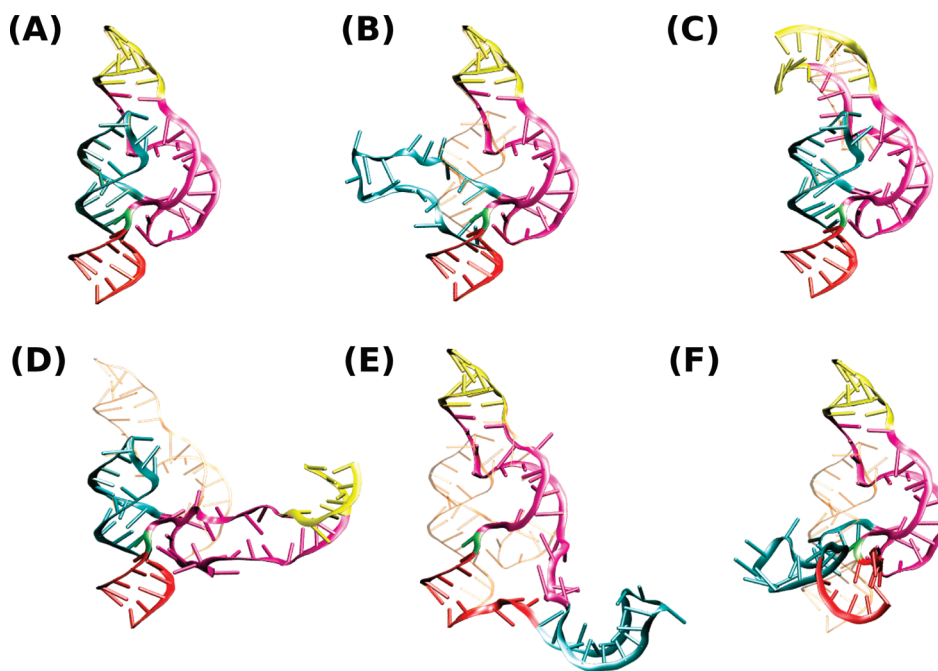


Figure 9. Sample structures from each cluster under the histogram in Figure 8. Structures are color coded using the same scheme as Figure 7. Each structure has been aligned for maximal overlap with the native conformation shown in translucent orange.

interchain loop closure move, but they involve hinges from the two different chains very close to the nucleolytic site, which is colored green, with one hinge on the uridine turn (purple, chain A) and the other hinge on stem III (red, chain B). This kind of loop motions leads to a complete unraveling of the catalytic site.

Since the histogram in Figure 8 was derived from an equilibrium sampling of the conformations of the molecule, the logarithm of the population should be proportional to the negative of the free energy divided by kT . Therefore, Figure 8 could be interpreted as a picture of the probability of the conformation as a function of rmsd away from the native state. However, we will refrain from taking this interpretation, since the potential energy function we have used is at best incomplete without a rigorous treatment of the counterions, and by using only two loop-closure MC move (3 + R + 3 MC1 and single-nucleotide MC3), the conformations sampled in the simulations may not have been completely equilibrated. Nonetheless, the results in Figure 8 show a close correspondence with the experimental results obtained by Liley et al.^{29–31} using fluorescence resonance energy transfer measurements, which show that in the absence of added salt the hammerhead unfolds into an extended three-way junction. This is consistent with the distribution in Figure 8 as well as some of the extended structures in Figure 9 identified by the simulations. This example has clearly demonstrated the feasibility of loop closure MC for investigating large-scale loop motions in RNA simulations.

7. CONCLUSION

In this paper, we have described the formulation of a loop closure problem that is applicable to multinucleotide loops of arbitrary lengths in RNAs. By representing the boundary constraints in the original loop closure problem using a new set of variables called the RETO coordinates, the original six-variable/six-constraint closure problem can be reduced to a simpler four-variable/four-constraint problem. This generalization permits a

simple solution of the multinucleotide loop closure problem. In various limits, this formulation is related to the conrot algorithm and the rebridging MC method. Using this generalized solution, we have developed new MC algorithms that are able to reclose loops of any arbitrary lengths to study large-scale loop motions in an all-atom RNA simulation. We have demonstrated the feasibility of the proposed method on the hammerhead ribozyme with a bound substrate and have shown that the simulation produced a large diversity of loop reconfigurations which were otherwise difficult to obtain from a conventional molecular dynamics or MC simulation.

AUTHOR INFORMATION

Corresponding Author

*To whom correspondence should be addressed E-mail: cmak@usc.edu.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under CHE-0713981.

REFERENCES

- (1) Go, N.; Scheraga, H. Ring closure and local conformational deformations of chain molecules. *Macromolecules* **1970**, *3*, 178.
- (2) Wakana, H.; Wako, H.; Saito, N. Monte-Carlo Study on Local and Small-Amplitude Conformational Fluctuation in Hen Egg-White Lysozyme. *Int. J. Pept. Prot. Res.* **1984**, *23*, 315.
- (3) Knapp, E. W. Long-Time Dynamics of a Polymer with Rigid Body Monomer Units Relating to a Protein Model - Comparison with the Rouse Model. *J. Comput. Chem.* **1992**, *13*, 793.
- (4) Dodd, L. R.; Boone, T. D.; Theodorou, D. N. A concerted rotation algorithm for atomistic Monte Carlo simulation of polymer melts and glasses. *Mol. Phys.* **1993**, *78*, 961.

- (5) Knapp, E. W.; Irgensdefregger, A. Off-Lattice Monte-Carlo Method with Constraints - Long-Time Dynamics of a Protein Model without Nonbonded Interactions. *J. Comput. Chem.* **1993**, *14*, 19.
- (6) Coutsiias, E. A.; Seok, C.; Jacobson, M. P.; Dill, K. A. A kinematic view of loop closure. *J. Comput. Chem.* **2004**, *25*, 510.
- (7) Brucoleri, R. E.; Karplus, M. Chain Closure with Bond Angle Variations. *Macromolecules* **1985**, *18*, 2767.
- (8) Sartori, F.; Melchers, B.; Bottcher, H.; Knapp, E. W. An energy function for dynamics simulations of polypeptides in torsion angle space. *J. Chem. Phys.* **1998**, *108*, 8264.
- (9) Deem, M. W.; Bader, J. S. A configurational bias Monte Carlo method for linear and cyclic peptides. *Mol. Phys.* **1996**, *87*, 1245.
- (10) Dinner, A. R. Local deformations of polymers with nonplanar rigid main-chain internal coordinates. *J. Comput. Chem.* **2000**, *21*, 1132.
- (11) Bashford, D.; Case, D. A. Generalized born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129.
- (12) Favrin, G.; Irback, A.; Sjunnesson, F. Monte Carlo update for chain molecules: Biased Gaussian steps in torsional space. *J. Chem. Phys.* **2001**, *114*, 8154.
- (13) Ulmschneider, J. P.; Jorgensen, W. L. Monte Carlo backbone sampling for polypeptides with variable bond angles and dihedral angles using concerted rotations and a Gaussian bias. *J. Chem. Phys.* **2003**, *118*, 4261.
- (14) Ulmschneider, J. P.; Jorgensen, W. L. Polypeptide folding using Monte Carlo sampling, concerted rotation, and continuum solvation. *J. Am. Chem. Soc.* **2004**, *126*, 1849.
- (15) Ulmschneider, J. O.; Jorgensen, W. L. Monte Carlo backbone sampling for nucleic acids using concerted rotations including variable bond angles. *J. Phys. Chem. B* **2004**, *108*, 16883.
- (16) Wu, M. G.; Deem, M. W. Efficient Monte Carlo methods for cyclic peptides. *Mol. Phys.* **1999**, *97*, 559.
- (17) Wu, M. G.; Deem, M. W. Analytical rebridging Monte Carlo: Application to cis/trans isomerization in proline-containing, cyclic peptides. *J. Chem. Phys.* **1999**, *111*, 6625.
- (18) Ho, B. K.; Coutsiias, E. A.; Seok, C.; Dill, K. A. The flexibility in the proline ring couples to the protein backbone. *Protein Sci.* **2005**, *14*, 1011.
- (19) Mak, C. H. RNA conformational sampling: 1. Single-nucleotide loop closure. *J. Comput. Chem.* **2008**, *29*, 926.
- (20) Lee, H.-Y.; Liang, C.-G. Displacement analysis of the general spatial 7-link 7R mechanism. *Mech. Mach. Theor.* **1988**, *23*, 219.
- (21) Kraulis, P. J. MOLSCRIPT: A Program to Produce Both Detailed and Schematic Plots of Protein Structures. *J. Appl. Crystallogr.* **1991**, *24*, 946.
- (22) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of the state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087.
- (23) Martick, M.; Scott, W. G. Tertiary contacts distant from the active site prime a ribozyme for catalysis. *Cell* **2006**, *126*, 309.
- (24) Humphrey, W.; Dalke, A.; Schulten, K. VMD - Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14*, 33.
- (25) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179.
- (26) 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.
- (27) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127.
- (28) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate radii. *J. Phys. Chem. A* **1997**, *101*, 3005.
- (29) Bassi, G. S.; Murchie, A. I.; Walter, F.; Clegg, R. M.; Lilley, D. M. Ion-induced folding of the hammerhead ribozyme: a fluorescence resonance energy transfer study. *EMBO J.* **1997**, *16*, 7481.
- (30) Hammann, C.; Lilley, D. M. Folding and activity of the hammerhead ribozyme. *ChemBioChem* **2002**, *3*, 690.
- (31) Penedo, J. C.; Wilson, T. J.; Jayasena, S. D.; Khvorova, A.; Lilley, D. M. Folding of the natural hammerhead ribozyme is enhanced by interaction of auxiliary elements. *RNA* **2004**, *10*, 880.

MSCALE: A General Utility for Multiscale Modeling

H. Lee Woodcock,^{*,†,⊥} Benjamin T. Miller,^{‡,⊥} Milan Hodoscek,[§] Asim Okur,[‡] Joseph D. Larkin,[‡] Jay W. Ponder,^{||} and Bernard R. Brooks^{*,‡}

[†]Department of Chemistry, University of South Florida, 4202 E. Fowler Avenue, CHE205, Tampa, Florida 33620-5250, United States

[‡]Laboratory of Computational Biology, National Heart Lung and Blood Institute, National Institutes of Health, Bethesda, Maryland 20892, United States

[§]Center for Molecular Modeling, National Institute of Chemistry, Hajdrihova 19, SI-1000 Ljubljana, Slovenia

^{||}Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, 660 S. Euclid Avenue, Box 8231, St. Louis, Missouri 63110, United States

S Supporting Information

ABSTRACT: The combination of theoretical models of macromolecules that exist at different spatial and temporal scales has become increasingly important for addressing complex biochemical problems. This work describes the extension of concurrent multiscale approaches, introduces a general framework for carrying out calculations, and describes its implementation into the CHARMM macromolecular modeling package. This functionality, termed MSCALE, generalizes both the additive and subtractive multiscale scheme [e.g., quantum mechanical/molecular mechanical (QM/MM) ONIOM-type] and extends its support to classical force fields, coarse grained modeling [e.g., elastic network model (ENM), Gaussian network model (GNM), etc.], and a mixture of them all. The MSCALE scheme is completely parallelized with each subsystem running as an independent but connected calculation. One of the most attractive features of MSCALE is the relative ease of implementation using the standard message passing interface communication protocol. This allows external access to the framework and facilitates the combination of functionality previously isolated in separate programs. This new facility is fully integrated with free energy perturbation methods, Hessian-based methods, and the use of periodicity and symmetry, which allows the calculation of accurate pressures. We demonstrate the utility of this new technique with four examples: (1) subtractive QM/MM and QM/QM calculations; (2) multiple force field alchemical free energy perturbation; (3) integration with the SANDER module of AMBER and the TINKER package to gain access to potentials not available in CHARMM; and (4) mixed resolution (i.e., coarse grain/all-atom) normal mode analysis. The potential of this new tool is clearly established, and in conclusion, an interesting mathematical problem is highlighted, and future improvements are proposed.

1. INTRODUCTION

Although computational methodologies have improved vastly over the last 10 years, it has become blatantly obvious that the most commonly employed techniques are not ideal for solving the challenging problems that exist at the interface of biology, chemistry, physics, and medicine. Many of the most important events surrounding biomedical processes take place on different time and length scales. For example, electronic excitations typically occur on the femto to picosecond time scale whereas aggregation, folding, and diffusion events can range in time from microseconds to hours.¹ This corresponds to approximately 10 orders of magnitude in the spatial regime and 15 orders of magnitude in the temporal regime.

In the past these multiple (time and length) scales have by and large been treated independently. The most notable exception is the coupling of quantum and classical (i.e., molecular mechanical) mechanics in a hybrid quantum mechanical/molecular mechanical (QM/MM) treatment. This scheme, which was first devised by Warshel and Levitt² with subsequent work by Singh and Kollman³ and Field, Bash, and Karplus,⁴ involves division of the system of interest into three regions. The first region is treated with quantum mechanics, while the larger second region

is described with MM. The third region is the smallest and describes the boundary between the QM and the MM sections. Inspired by the success of this methodology, Morokuma and co-workers introduced a general approach to coupling different levels of theory dubbed the ONIOM method.⁵

The term multiscale modeling typically describes the use of disparate methods to solve problems that span methodological, temporal, or spatial scales. For example, hybrid QM/MM schemes utilize two different methodological scales. It is generally accepted that there are two main approaches used in multiscale modeling: sequential and concurrent.^{6,7} The sequential multiscale treatment employs more accurate models that are then used to parametrize coarser ones. Coarse grained (CG) molecular dynamics (MD) simulations is one area where the idea of sequential multiscale modeling has been particularly useful. In general, coarse graining seeks to accurately represent a system with a reduced number of degrees of freedom. Examples of this include treating an atomistic amino acid residue as a single or series of beads (e.g., BLN model) or representing an α carbon as

Received: December 22, 2010

Published: March 25, 2011

an elastic or Gaussian network [e.g., elastic network model (ENM), Gaussian network model (GNM), etc.].^{8–24} In these cases the sequential, force-matching approaches can significantly improve results as they typically use classical atomistic simulations to derive coarse grain parameters that ideally reproduce the desired properties of the parent system.^{25–27} The other main approach is concurrent multiscale modeling which is exemplified by the hybrid QM/MM scheme; instead of using one model to improve another, both models are executed simultaneously on different parts of the system. Further, concurrent modeling can be subdivided into additive and subtractive approaches with the contrasting ways used to couple QM and MM methodologies perfectly highlighting the two subcategories.

The original QM/MM scheme from Warshel and Levitt is an additive concurrent approach where the interactions that couple the QM region to the MM region are comprised of a hybrid Hamiltonian (including polarization effects from the MM regions) that are “added” to the total energy of the system:

$$\hat{H}_{\text{eff}} = \hat{H}_{\text{QM}} + \hat{H}_{\text{MM}} + \hat{H}_{\text{QM/MM}} \quad (1)$$

where \hat{H}_{QM} is the pure QM Hamiltonian, \hat{H}_{MM} is the classical Hamiltonian, and $\hat{H}_{\text{QM/MM}}$ is the hybrid QM/MM Hamiltonian. In contrast, Morokuma and co-workers developed a subtractive approach where the interaction between the QM and MM systems is described only at the MM level of theory:

$$E^{\text{ONIOM}} = E_{\text{real}}^{\text{MM}} + E_{\text{model}}^{\text{QM}} - E_{\text{model}}^{\text{MM}} \quad (2)$$

where the final E^{ONIOM} is the final extrapolated energy and is meant to approximate the full system treated at the QM level of theory. The first term on the right-hand side describes the “real” system, which is comprised of all the atoms (treated at the lowest level of theory, MM). The second term on the right-hand side is the energy of the model system (i.e., the region treated at the highest level of theory, QM), while the final term repeats the model system calculation, however, only treated at the low level. This final term is needed to prevent double counting of the model system. Note the fundamental difference between the additive and subtractive methodologies here; the final term in eq 1 describes the interactions between the QM and MM regions directly, however, in eq 2 this interaction is completely encompassed as part of the first term ($E_{\text{real}}^{\text{MM}}$). The subtractive scheme is sometimes referred to as “mechanical embedding.” It should be noted that this incarnation of the ONIOM scheme does not accommodate direct polarization of the model region by the remaining portion of the system, however, improvements to this scheme have been developed to overcome this weakness.²⁸

In general, it is believed that the additive approach is more robust and accurate, however, that is predicated on deriving and implementing the hybrid coupling term which can in practice be rather difficult.^{29,30} Therefore, the subtractive approach has the attractive features of being both conceptually easy to understand and implement. Additionally, further development has been carried out to extend the subtractive QM/MM approach with ideas from the additive scheme (i.e., electronic embedding).²⁸

Although QM/MM is clearly the most widely utilized multiscale method, the modeling community continues to clamor for more general approaches. A recent review correctly highlights this point: “very few packages implement the CG and coarser models, and even fewer integrate more levels in a fully multiscale software. An effort in this direction would be very useful and would promote the use of multiscale approaches.”⁶

In the following sections we describe such a general framework, called MSCALE. We briefly review the conceptual basis of our current multiscale approach and describe the implementation of this scheme within CHARMM.³¹ Periodic systems and symmetry are fully supported within this framework. The Results and Discussion Section highlights four representative examples that showcase a range of the functionality supported by MSCALE. The first example will highlight the use of MSCALE as a general subtractive, either QM/MM or QM/QM, engine. The second example will demonstrate the ability to combine multiple classical force fields into a single calculation and interface into the alchemical free energy perturbation module of CHARMM. A third case will describe the implementation of AMBER’s SANDER module and the TINKER software suite as “servers” to CHARMM’s MSCALE command. This functionality allows CHARMM users a general way to access features exclusive to AMBER or TINKER (e.g., implicit water models, polarizable force fields) and can easily be extended to support new developments in both programs. The final example will showcase the ability to perform a mixed model normal mode analysis (NMA) by combining both atomistic and CG treatments.

2. COMPUTATIONAL DETAILS

The current work describes the extension of concurrent multiscale approaches and introduces a general framework called MSCALE to access this functionality, which has been fully implemented in CHARMM via the MSCALE command. Throughout this paper, MSCALE will be used to refer to the general framework, while MSCALe refers to the specific CHARMM command. This multiscale approach makes the coupling of both additive and subtractive schemes possible within a single framework and allows for both types of methods to be used in a single calculation.

To actually perform an MSCALE calculation, the MSCALE command is invoked by the user, followed by one or more SUBSystem commands to define the different structural regions of the calculation. An executable must be given along with an input file, output file path, weighting coefficient of the subsystem, other optional parameters, and an atom selection (i.e., atomic coordinates) as arguments to the SUBSystem command. The executable may be any MSCALE-compatible program (although not all features are supported on codes other than CHARMM). Examples of optional arguments to SUBSystem are the CRYStal keyword to specify that periodic image data should be transmitted and the NPROC argument, which specifies how many processors the subsystem should use in a parallel/parallel calculation. Example input scripts are given in the Supporting Information.

Once all subsystems are defined, the user can perform calculations as usual. The calculation is executed in a parallel client/server fashion (Figure 1), with basic communication being handled by version 2 of the standard message passing interface (i.e., MPI-2).³² The “client” acts as the controlling process with “server” calculations being spawned based on the SUBSystem commands entered by the user; one server is launched for each defined subsystem using the executable, atom selection, and other parameters specified by the user. This spawning is done through the MPI_Comm_spawn MPI-2 routine. The client stores the MPI intercommunicator needed to transfer data to and from the newly spawned server. Whenever the energy, gradient, or Hessian is required, the client sends each server the coordinates it needs and, if necessary, other data such as unit cell dimensions for a periodic system via an MPI broadcast on that server’s intercommunicator. It then waits to receive the energy,

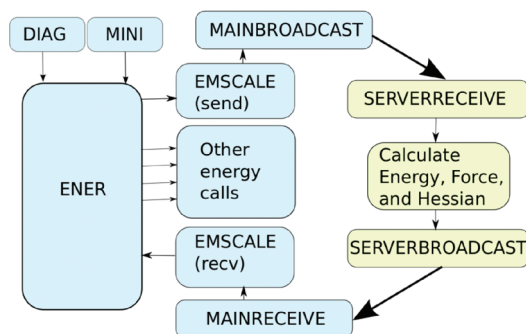


Figure 1. Illustration of the subroutine calling sequence of the MSCALE facility, showing the information flow of a typical energy (ENER), minimization (MINI), or normal mode analysis (DIAG) calculation. The broadcast and receive routines handle both coordinate and energy/gradient communication. Routines in blue are executed on the main processor (client), while those in yellow take place on the subsystems (servers). Thin black lines represent information being passed between subroutines where the thick black lines represent MPI calls and the sharing of information between the controlling client process and the server process, which acts only as an energy, force, or Hessian engine. The EMSCALE subroutine is called twice from CHARMM's main energy routine, once at the beginning to send the data to the servers and again at the end to receive the energy, force, etc. terms from them. Therefore, the servers and the clients are performing calculations in parallel. Further details of how MSCALE is implemented is given in the Supporting Information.

gradient, and optional Hessian elements (the virial is also transmitted for periodic systems) from each server, scales them by the specified subsystem coefficient, and adds them to those terms calculated by the client. If it is not desired to include the terms calculated on the client in the final energy, CHARMM's SKIPE or BLOCk commands can be used to discard them.

Currently, only CHARMM³¹ is supported as the client with the server calculations being able to employ either CHARMM, TINKER,^{33,34} or the SANDER module of AMBER³⁵ as servers directly through MPI calls. Periodic systems are not fully supported for non-CHARMM servers, but this functionality will be added in the future. There is an additional set of quantum chemical programs that are supported as servers. These packages are interfaced through external wrappers that handle MPI communication. At this time the following QM packages are supported this way: NWChem,³⁶ Molpro,³⁷ PSI3,³⁸ and Gaussian 03/09.³⁹ It should be noted that all of CHARMM's "built-in" QM packages (e.g., Q-Chem,⁴⁰ GAMESS-US,⁴¹ GAMESS-UK,⁴² SCC-DFTB,⁴³ MNDO,⁴⁴ QUANTUM,⁴ and SQUANTUM,⁴⁵ etc.) are supported directly via MPI. Using this functionality both additive and subtractive methodologies can be combined in a single calculation.

In a mixed scale calculation, coarse grained centers can be handled by either colocating them with a center from the all-atom model (e.g., a C_α) or by connecting them through constraints or restraints. For example, CHARMM's LONEpair facility can be employed to constrain a single CG center to be located at the center of mass of a group of atoms.³¹ Using one of these two options we have defined a general procedure for mixing models of different resolution.

In this manner, MSCALE supports multiple independent, but connected, calculations. The user is free to define an unlimited number of subsystems (i.e., layers) and assign arbitrary coefficients to combine them. The individual calculations are run as separate processes, usually on different computers (e.g., on a Beowulf style

cluster).⁴⁶ For example, the typical subtractive QM/MM approach, (i.e., executed as an ONIOM-type calculation) would have a coefficient matrix as such:

	Real	model
MM	1	-1
QM	0	1

with the final energy expression being that of eq 2. Using the MSCALE utility, this general approach can be extended to arbitrary levels of theory (i.e., QM, MM, coarse grain, etc.) and arbitrary numbers of subsystems. An example of a four-layer subsystem matrix:

	Full	Big	Medium	Small
\$	1	-1	0	0
\$\$	0	1	-1	0
\$\$\$	0	0	1	-1
\$\$\$\$	0	0	0	1

where the left side of the matrix represents the cost of level of theory (increasing from top to bottom), and the top represents the size of the system (increasing from right to left). This matrix yields the following MSCALE energy:

$$E^{\text{MSCALE}} = E_{\text{S}}^{\text{Full}} + (E_{\text{SS}}^{\text{Big}} - E_{\text{S}}^{\text{Big}}) + (E_{\text{SSS}}^{\text{Medium}} - E_{\text{SS}}^{\text{Medium}}) + (E_{\text{SSSS}}^{\text{Small}} - E_{\text{SSS}}^{\text{Small}}) \quad (3)$$

Using this equation, one can easily derive the force and the Hessian expressions with the proper link-atom projections.^{47,48} More complex multiscale systems can be set up that involve a combination of additive and subtractive schemes. This can be done easily, without any need for reprogramming, and there is effectively no limit to the number of different subsystems.

The coefficients on the subsystems are not just used for ONIOM-type calculations. They may also change dynamically based on the λ values used in CHARMM's alchemical free energy PERTurbation procedure.³¹ This method calculates the free energy difference (ΔG) between two molecular systems. To do so, an initial system (the $\lambda = 0$ state) is defined, and then some change (e.g., changing the protonation state of a titratable amino acid) is made to define a second state (the $\lambda = 1$ state). A MD simulation is then run during which λ is gradually moved from 0 to 1. The free energy difference between these two end states may then be estimated by thermodynamic integration⁴⁹ (thermodynamic perturbation is also supported). With MSCALE, subsystems may be given a weighting of λ or $1 - \lambda$, which allows the contribution of these subsystems to the total energy to change over the course of the simulation. When energy terms are scaled by λ or $1 - \lambda$, derivatives of the energy and forces with respect to λ are computed and applied appropriately. This allows MSCALE to be fully compatible with other free energy methods in CHARMM. Compatibility of MSCALE and PERT is important because until now the PERT facility has been limited in functionality, and in many cases, the direct implementation of new methods in PERT was both conceptually and technically challenging. Now, using MSCALE individual CHARMM (or other programs as described above) processes can be spawned that are not dependent on the limited PERT module. One example of this is the use of additive QM/MM; it was not until recently that PERT supported an ab initio QM program.⁵⁰ However, with MSCALE all currently supported quantum packages in

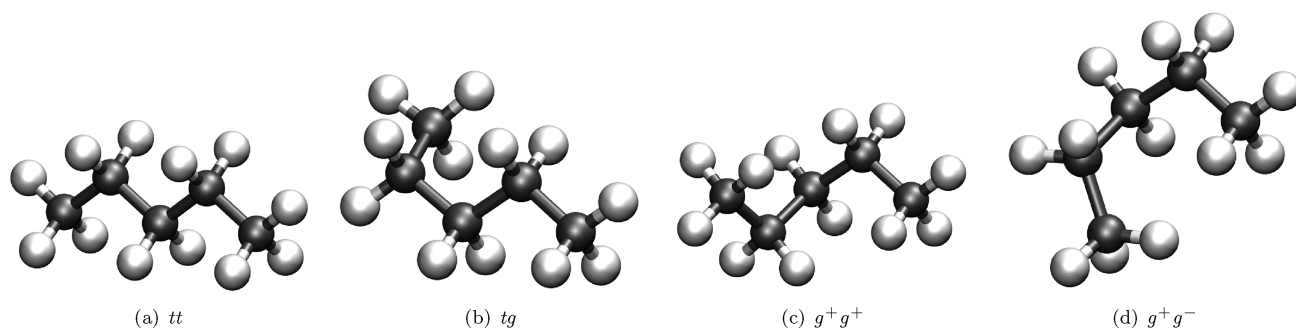


Figure 2. Conformations of pentane.

CHARMM will have access to free energy perturbation. Another example is that, until this development, the only alchemical changes that could be studied with PERT are those that could be represented by a structural change (e.g., changing the type or connectivity of an atom or group of atoms). Because each subsystem is entirely independent under MSCALE, non-structure based changes (e.g., changing force field parameters or even the entire force field) can be made. This is demonstrated below with an example of estimating the change of ΔG of solvation in moving a system from the CHARMM22 force field⁵¹ to the AMBER99SB force field.⁵²

Initially, one thinks about mixing atomistic and CG models, but this facility will allow the mixing of multiple CG models as well. An example of this would be the combination of a 1 site ENM with a 1 or multisite beaded CG model. An interesting application of this could be to study protein–protein interactions where the beaded model can better describe local interactions, while the ENM ensures proper protein structure. Additionally, two atomistic models can be combined just as easily. For example, there is great need in the biomedical engineering community to be able to model protein–surface interactions with a consistent approach.^{53–56} This facility allows for using a proper materials-based potential while combining that with state-of-the-art biological force fields, like CHARMM, AMBER, OPLS, etc.

The MSCALE framework and its implementation in CHARMM fully support periodic systems, as mentioned above. The accurate single-sum pressure calculation is used.^{31,57} This single-sum method can be used in CHARMM without MSCALE support, since image atom positions are explicitly treated and used when computing the internal virial. This avoids objections to this methodology that have been raised in previous work.⁵⁸ Within CHARMM's MSCALE implementation, there is support for constant pressure and temperature simulations even when the “servers” do not support virial calculations.

As alluded to above, MSCALE currently supports analytic and finite difference energies, forces, and Hessians (Currently, only Q-Chem supports QM and QM/MM Hessians). Therefore, minimization, dynamics, and normal mode analysis are all possible within this framework. In the following section several examples of this functionality will be presented. It should be stressed that the goal of this effort is to provide an extensive and flexible tool for multiscale modeling, as opposed to a single tool with a dominant focus on performance. The use of multiple processes allows for simple parallelization that can enhance performance, but communicating selected coordinates and associated energies, forces, and optionally Hessian terms on every step does entail a performance cost. CHARMM's MSCALE

implementation does allow subsystems to be run on multiple processors if the server program has been parallelized.

3. RESULTS AND DISCUSSION

3.1. Subtractive QM/MM and QM/QM. Subtractive MSCALE calculations of pentane in its trans–trans (*tt*), trans–gauche (*tg*), gauche plus–gauche plus (g^+g^+), and gauche plus–gauche minus (g^+g^-) conformations (Figure 2) were performed. All QM/MM and QM/QM calculations utilized the Q-Chem/CHARMM⁵⁰ or SCC-DFTB/CHARMM⁴³ interfaces. All “high-level” QM results were obtained at the MP2/cc-pVDZ level of theory, except for the full-QM results for which MP2/cc-pVTZ was used. The CHARMM force field was used for MM and the semiempirical SCC-DFTB method employed as the “low-level” QM in QM/QM calculations. The MP2/cc-pVDZ level was chosen as to be comparable to previously published work.⁵⁹ All subtractive calculations were repeated four times with 1, 2, 3, and 4 methyl groups included as part of the QM region (Figure 2). The single link atom (SLA) boundary scheme coupled with the LONEpair facility, to ensure colinearity of the link atom, was used for all additive QM/MM calculations. Results were obtained by performing 50 steps of steepest descent minimization followed by adopted basis Newton Raphson (ABNR) minimization until reaching a root mean squared gradient (GRMS) tolerance of 0.001 kcal/mol \AA^2 except for the full QM case, which was minimized in Q-Chem to a gradient tolerance of at least 1.5×10^{-5} hartree/bohr using the MP2/cc-pVDZ level of theory. Full QM, additive QM/MM, subtractive QM/MM, and subtractive QM/QM energy differences of all conformations are reported in Table 1. In all cases the *tt* conformation was the global minimum with the *tg*, g^+g^+ , and g^+g^- energy differences progressively increasing.

One area of interest that can be addressed from these results centers around the recent hypothesis of adjacent gauche stabilization.⁵⁹ Klauda et al. found that adding a single gauche state to an alkane resulted in a 0.54–0.62 kcal/mol penalty. Further, they reported adding a second gauche state of the same sign required 0.22–0.37 kcal/mol of energy while adding one of opposite sign cost 2.49–2.85 kcal/mol. Examining the ΔE s from Table 1 reveals several trends.

Beginning with the *tt*–*tg* conformational change it is clear from all results that treating only the terminal methyl group classically overestimates the ΔE (~ 0.68 kcal/mol). This is in part due to the neglect of long-range dispersive interactions between the terminal methyl groups which occurs in all three models. Specifically, this is a result of the energy differences of the MP2 subsystem ($\text{CH}_3\text{--CH}_2\text{--CH}_2\text{--CH}_3$), 0.67 and 0.65 kcal/mol

Table 1. Energy Differences in kcal/mol between the Optimized Trans–Trans Conformation (*tt*) and the Trans–Gauche (*tg*), Positive Gauche–Positive Gauche (g^+g^+), and Positive–Gauche–Negative Gauche (g^+g^-) Conformations of Pentane^a

<i>N</i>	<i>tg</i>	g^+g^+	g^+g^-
Additive QM/MM			
1 (CH ₄)	0.62	1.07	2.79
2 (CH ₃ CH ₃)	0.28	0.82	2.59
3 (CH ₃ CH ₂ CH ₃)	0.48	0.61	2.73
4 (CH ₃ CH ₂ CH ₂ CH ₃)	0.68	0.96	2.95
Subtractive QM/MM			
1 (CH ₄)	0.60	1.13	2.83
2 (CH ₃ CH ₃)	0.56	1.10	2.78
3 (CH ₃ CH ₂ CH ₃)	0.55	1.05	2.97
4 (CH ₃ CH ₂ CH ₂ CH ₃)	0.68	1.22	3.04
Subtractive QM/QM			
1 (CH ₄)	0.47	0.92	2.32
2 (CH ₃ CH ₃)	0.43	0.89	2.26
3 (CH ₃ CH ₂ CH ₃)	0.53	0.95	2.36
4 (CH ₃ CH ₂ CH ₂ CH ₃)	0.68	1.14	2.66
full QM	0.56	0.80	2.81
full SCC-DFTB	0.47	0.92	2.32
full MM	0.60	1.13	2.82

^aThe leftmost column (*N*) indicates the number of groups represented at the MP2/cc-pVDZ level of theory. QM/QM refers to mixed MP2/cc-pVDZ - SCC-DFTB calculations.

for the subtractive QM/MM and QM/QM, respectively. The additive QM/MM case mirrors this since the electrostatics on the terminal methyl group are excluded (vide infra).

The subtractive results, going from two to four methyl groups, converge relatively smoothly toward the results of the low level of theory (i.e., MM and SCC-DFTB, respectively). However, there is significant variation in the additive QM/MM results. This is easily explained as an artifact of the SLA approach coupled with excluding group electrostatics (EXGR), which is a standard scheme for preventing over polarization in QM/MM calculations. Although this clearly leads to underestimation of energy differences, it is fairly consistent throughout all conformations. Further, Das et al. showed this effect can be mitigated easily by using more advanced link atom approaches (e.g., delocalized Gaussian MM charges).⁶⁰

Next, examining the $tt-g^+g^+$ ΔE s leads again to clear trends. For example, the subtractive QM/MM calculations fail to reproduce the adjacent gauche penalty (0.22–0.37 kcal/mol). This results from a combination of both the neglect of dispersive effects and overestimation by the underlying force field, Table 2. Due to the correct handling of dispersion at the MP2 level of theory and explicit polarization, it is clear why additive QM/MM does a better job of reproducing this subtle effect. The other subtle effect at play here is the adjacent gauche destabilization which results in an ~ 2.81 kcal/mol ΔE between *tt* and g^+g^- . Again, additive QM/MM handles this interaction relatively well as does subtractive QM/MM; largely due to parametrization of the CHARMM force field, ΔE ($tt-g^+g^-$) = 2.82 kcal/mol. However, the problem with incorrect treatment of dispersion pops up again; SCC-DFTB ΔE ($tt-g^+g^-$) underestimates this by nearly 0.5 kcal/mol.

Overall, the results indicate that the additive and subtractive QM/MM and QM/QM methods both do a reasonable job of reproducing the energy surface for alkane molecules at a substantially lower computational cost. However, specific choices of

Table 2. Conformational Energy Differences Between the *tg* and g^+g^+ States.^a

level of theory	ΔE
MP2/cc-pVDZ	0.24
SCC-DFTB	0.45
CHARMM	0.53

^aThese highlight the effect of level of theory on adjacent gauche stabilization.

model systems and levels of theory can lead to incorrect descriptions of subtle effects. Use of the subtractive method with MSCALe allowed for combined quantum mechanical and semiempirical calculations, which are not possible with previously existing methods in CHARMM.

3.2. Free-Energy Perturbation. The PERT module in CHARMM is a single topology-based potential implemented to calculate alchemical and/or conformational free energies. CHARMM's MSCALe implementation has been integrated with this module and vastly expands the features compatible with PERT. All MSCALe + PERT simulations were run without SHAKE constraints unless otherwise noted, but the use of SHAKE with MSCALe is supported. Basic PERT functionality was tested with MSCALe using two simple systems. The first examined migration of the hydroxyl group (–OH) of methanol (CH₃OH) using the CHARMM22 force field.⁵¹ For the PERT run, 16.4 ns (ns) of vacuum MD was carried out using a 1 fs time step. The alchemical process was divided into 41 windows, each running 400 picoseconds (ps) with the first 200 ps being used for equilibration and final 200 used for collecting statistics. Lambda (λ) was increased by 0.025 at each window; this yielded a net free energy change (ΔG) of effectively 0, as would be expected. The ΔG as a function of PERT window is shown in Figure 3A.

The second simple test was reversing the chirality of a single alanine molecule ($R \rightarrow S$). For this simulation, 82 ns of Langevin dynamics (LD) was run with a collision frequency of 2 ps⁻¹; again utilizing 41 windows ($\Delta\lambda = 0.025$) with 2 ns of dynamics per window (1 ns for equilibration and 1 ns for data collection). Once again, the total free energy change was effectively 0, as would be expected. For validation the calculation was repeated with two different random seeds, and all simulations yielded ΔG values within a few hundredths of a kcal/mol from 0.

To extend this functionality to a more novel application, we examined the alchemical free energy of an alanine dipeptide moving from the CHARMM22 to AMBER (AMBER99SB) force field (as implemented in CHARMM).^{31,51,52} This was carried out in both vacuum and solvent using AMBER's version of the TIP3P water model. A consistent water model was used for both force fields to prevent solvent free energy changes from dominating the total ΔG . For both force fields, nonbonded interactions were calculated in full applying no cutoffs.

To validate the methodology, the vacuum structure was minimized to the C7 equatorial conformation, and a free energy perturbation calculation, changing the CHARMM22 to the AMBER99SB force field, was carried out. For the dynamics, LD was run at 0 K for 168 ns with a Langevin collision frequency of 1 ps⁻¹. The simulation consisted of 21 windows running 8 ns each with only the last 7 ns being used for data collection. Cut-offs were disabled by setting the nonbonded cutoff to 996 Å. Under these conditions, the measured $\Delta G_{\text{CHARMM} \rightarrow \text{AMBER}}$ was –6.15 kcal/mol, which is effectively identical to the difference in potential energy between the CHARMM and AMBER99SB force fields for this conformation,

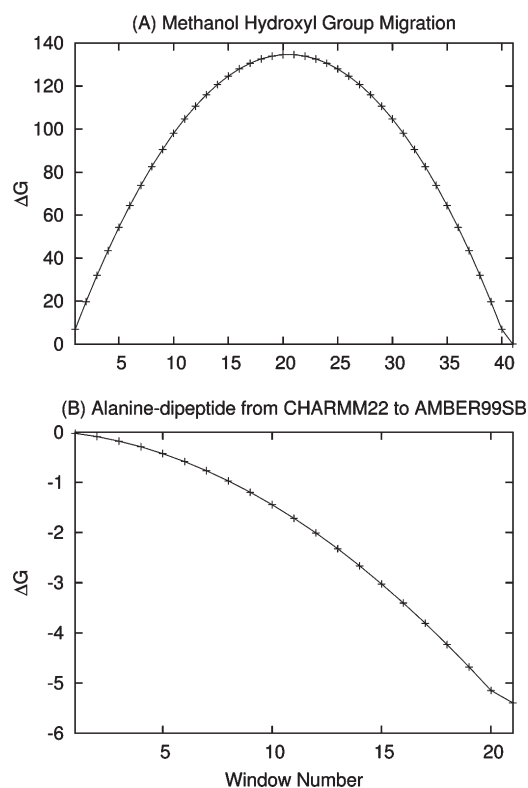


Figure 3. ΔG as a function of window for: (A) the OH move of methanol and (B) the alanine dipeptide moving from the CHARMM22 to AMBER99SB force fields. In both cases the free energy curve is smooth, representing a gradual shift from one force field or structure to another.

as would be expected with 0 K dynamics where entropic effects are nonexistent.

The vacuum structure was also run under the same PERT setup with LD at 300 K and a collision frequency of 1 ps^{-1} . In this case $\Delta G_{\text{CHARMM} \rightarrow \text{AMBER}}$ was -5.39 kcal/mol (ΔG as a function of window for this run is shown in Figure 3B). At 300 K, the structure samples multiple wells therefore a harmonic limit free energy is not expected to perfectly reproduce the dynamics. However, the harmonic limit $\Delta G_{\text{CHARMM} \rightarrow \text{AMBER}}$ was determined from normal mode analysis to be -5.44 kcal/mol for C7 equatorial conformation, -5.50 kcal/mol for the C7 axial conformation, and -5.10 kcal/mol for the C5 conformation. These are consistent with the ΔG obtained from the PERT simulation.

To explore the effects of solvent, the vacuum minimized C7 equatorial conformation was placed in a pre-equilibrated water sphere with a radius of 15 \AA (periodic boundary conditions were not used). LD was run at 300 K for 2.1 ns with a collision frequency of 0.5 ps^{-1} using the same windowing scheme as in the vacuum case but with each window containing only 100 ps of dynamics, of which the last 75 ps are used for collecting statistics. As in the case of the gas-phase systems, no cutoffs were applied. SHAKE was applied to the TIP3P waters only. Under these conditions, $\Delta G_{\text{CHARMM} \rightarrow \text{AMBER}}$ was calculated to be -6.48 kcal/mol .

The differences between current results and those reported by Boresch and co-workers⁶¹ can be explained by the fact we used CHARMM's AMBER99SB implementation instead of PARM94. Additionally, our solvated simulations were performed without

periodic boundary conditions and with no cutoffs; finally, our results are computed with thermodynamic integration as opposed to the Bennett Acceptance Ratio.⁶² The total change of free energy caused by solvation ($\Delta \Delta G_{\text{CHARMM} \rightarrow \text{AMBER}}^{\text{solvation}}$) under these conditions is -1.09 kcal/mol at 300 K. Much of the difference between this result and the previous one is likely explained by changes between the AMBER94 and AMBER99SB force fields and by the fact that the boundary conditions for the solvated system were different, as the previous work notes that they were able to obtain thermodynamic integration results consistent with those found with the Bennett acceptance ratio. Finally, the previous work notes that the choice of cutoff radius has a substantial effect on the calculated ΔG , and therefore applying a cutoff to this system is likely to yield a somewhat different $\Delta \Delta G_{\text{CHARMM} \rightarrow \text{AMBER}}^{\text{solvation}}$.

3.3. Integration with AMBER and TINKER. Through the MSCALE implementation in CHARMM, the SANDER module in AMBER³⁵ can be called to perform energy and analytic gradient calculations. Such an implementation enables external use of the AMBER energy function, AMBER force fields, and implicit solvation models while the controlling dynamics and/or analysis is carried out in CHARMM. Likewise, an interface was developed to the TINKER suite of programs,^{63,64} allowing access to the polarizable AMOEBA force field.^{33,34} Clearly the ability to interface with already developed simulation codes will save countless hours of reimplementing and validation of methods that are being ported from package to package.

3.3.1. AMBER Implementation. In order to implement the MSCALE communication paradigm in the SANDER module of AMBER, a new command line option (-server) was added to tell SANDER that instead of running an energy minimization or MD simulation, it should call a special routine that handles MSCALE communication. This routine waits for an MPI message from the master processor (the client) containing the coordinates of the subsystem it has been assigned to calculate. SANDER then calculates the energy and forces as normal using these coordinates and passes these back to the main processor (which is assumed to be running CHARMM). Essentially, the server-side routines (denoted by yellow boxes in Figure 1) were added to SANDER, with necessary changes (e.g., calling SANDER's energy routine instead of CHARMM's) being made. In addition to these changes to SANDER, a few small modifications needed to be made to CHARMM's MSCALE implementation. A new option (AMBER) was added to the SUBSystem command alerting CHARMM that the executable being called is AMBER's SANDER program and that the -server flag should be used as an argument to the specified executable.

To test the implementation, alanine dipeptide simulations were run in standard AMBER and the CHARMM–MSCALE–AMBER combination. In both cases the AMBER99SB force field⁵² and generalized Born (GB^{OC}) implicit solvent model⁶⁵ were used. LD with a 1 fs step and collision frequency of 1 ps^{-1} were used to include solvent friction and temperature control. All bonds involving hydrogen atoms were constrained using the SHAKE^{66,67} algorithm. Simulations were started from a linear conformation and run for 500 ns.

Analysis was performed on snapshots taken at 1 ps intervals. The first 10 ns (10 000 snapshots) were discarded during analysis for equilibration. Two-dimensional histograms of the dihedral angles were calculated with a bin size of $5^\circ \times 5^\circ$, and free energies were plotted using the populations of each bin by setting the most populated bin at 0 kcal/mol (Figure 4) using the matplotlib

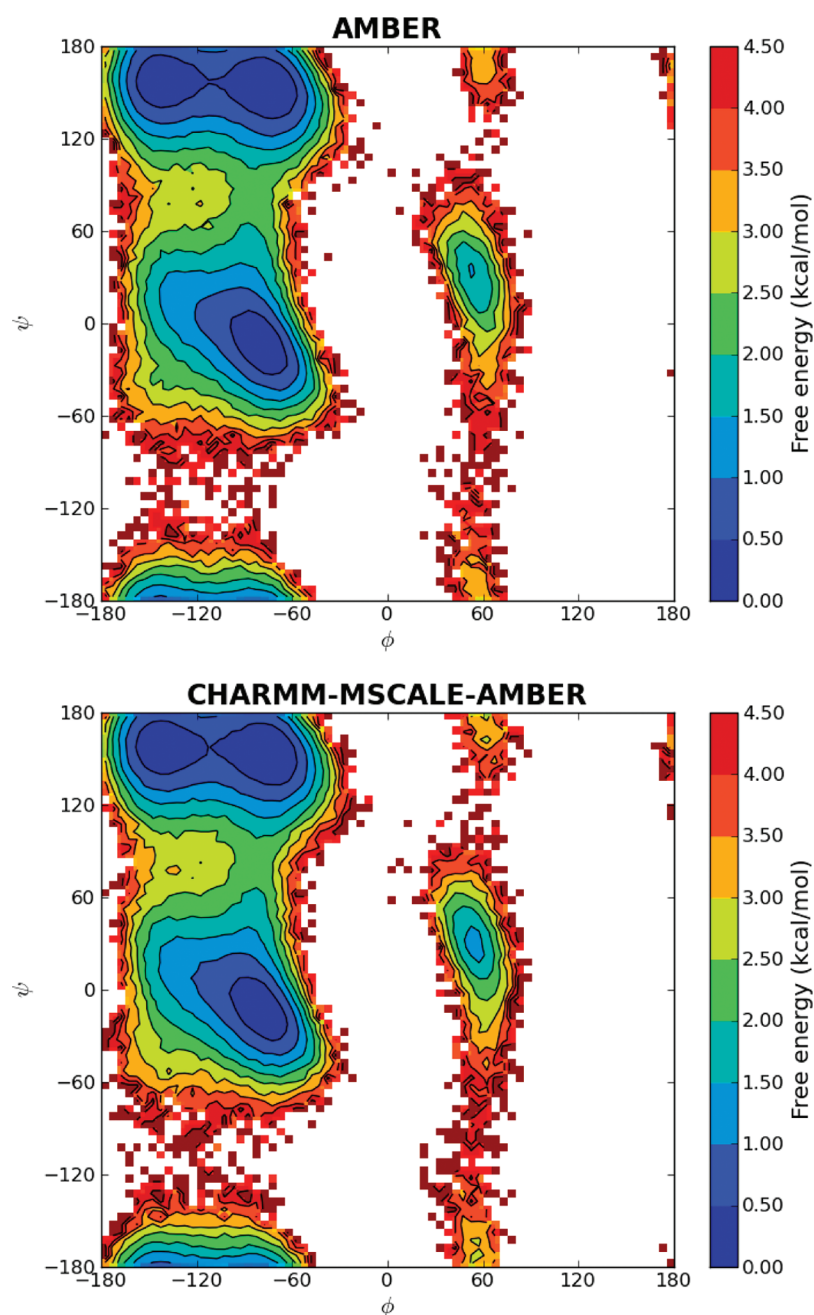


Figure 4. Ramachandran free energy landscapes of alanine dipeptide with standard AMBER and CHARMM–MSCALE–AMBER simulations. Both free energy profiles are very similar. Slight differences are expected even though the same force field and solvent method are used, since the MD runs were performed with different packages with their own implementations of LD.

module in Python version 2.6. As seen from Figure 4, almost identical free energy profiles and barriers were obtained for all relevant conformations (α , β , P^H and α_L). The minimum for the α_L region has the same free energy for both methods, but the MSCALE plot seems to be slightly broader. This is most likely a sampling issue since the α_L region is visited less frequently than the other dominant conformations.

3.3.2. TINKER Implementation. For the TINKER implementation of the MSCALE communication routines, a new program called *mslave* was written and linked against the main TINKER library. The only purpose of this program was to manage the communication between the “client” program and TINKER

running as a “server”. The functionality is similar to that enabled by the *-server* option to SANDER that is mentioned above, but because TINKER is implemented as a collection of programs rather than a monolithic binary, it is desirable to have a dedicated program to handle communication. The *mslave* program uses MPI to receive the coordinates from the client and calls the main TINKER energy and gradient routines and, if desired by the user, the appropriate routines to compute the Hessian. The energy, forces, and (if specified) Hessian are then returned to the client.

As a test case, the alanine dipeptide was again used. Six local minima, as defined by the AMOEBA force field, were located using the SEARCH program in TINKER. These conformations

Table 3. Initial Conformations (Obtained Using TINKER's SEARCH Program) and Final (i.e., Minimized) Energy Differences of the Alanine Dipeptide^a

conformation	ϕ	ψ	ΔE (AMOEBA)	ΔG (AMOEBA)	ΔE (C22)	ΔG (C22)
C7eq	-83.1	76.4	0.00 (-26.97)	0.00 (24.95)	0.00 (-14.10)	0.00 (37.92)
C7ax	72.1	-53.0	2.47 (-24.50)	2.40 (27.35)	2.08 (-12.02)	2.39 (40.31)
C5	-155.1	-162.6	1.20 (-25.77)	0.24 (25.19)	1.13 (-12.97)	0.35 (38.27)
B2	-116.8	10.8	2.78 (-24.19)	1.64 (26.59)	0.00 (-14.10)	0.00 (37.92)
aP	66.1	30.2	4.41 (-22.56)	4.27 (29.22)	2.08 (-12.02)	2.39 (40.31)
aL	-168.2	-35.2	5.54 (-21.43)	5.55 (30.50)	1.13 (-12.97)	0.35 (38.27)

^a Raw and free energies are listed in parentheses. The AMOEBA energies were calculated using the original starting conformations, whose ϕ and ψ angles are given. A minimization was performed with the CHARMM22 force field from each starting conformation. The resulting minimized structure was used to calculate the CHARMM22 (C22) energy and the free energy. All energies and free energies are in kcal/mol.

included the C7 equatorial, C7 axial, and C5 conformations used for the PERT test case above (although the AMOEBA minima were at slightly different positions than those found for the CHARMM22 force field). Additionally, three more local minima, denoted B2, aP, and aL, were found. The energies of these structures were evaluated using both TINKER directly and TINKER called from CHARMM through the MSCALE command to ensure equivalent results. Using TINKER through MSCALE, the Hessians were then calculated, and ΔG was determined in the harmonic limit. The values of the ϕ and ψ angles for the results are reported in Table 3. The six TINKER local minima were also each minimized using the CHARMM22 force field with default nonbonded cut-offs; energies and free energies of the resulting CHARMM22 local minima are also reported in Table 3.

The energies and free energies shown in Table 3 clearly illustrate the difference between the CHARMM22 and AMOEBA force fields. Of particular interest was the fact that the B2, aP, and aL minima from AMOEBA were not near any local minima under the CHARMM22 force field. When minimized using CHARMM22, these conformations fell back to the C7eq, C7ax, and C5 conformations, respectively. It is not surprising that the AMOEBA force field presents a rougher energy surface with more local minima, since it is a polarizable force field that is more complex than CHARMM22.

3.4. Multiscale Normal Mode Analysis. In addition to supporting the multiscale evaluation of energies and forces, the MSCALE facility also fully supports analytic and finite difference Hessian calculations (i.e., normal mode analysis). In the same way that energies and forces are returned from the server process to the controlling client process, Hessian matrix elements are also passed back and mapped to their full system counterparts. In this way both fully atomistic and mixed model normal mode analyses can be performed. Of particular interest is the ability to combine coarse grained models with all-atom representations, such that the most interesting part of a macromolecular system (e.g., a binding domain) can be represented as all-atom, while the larger scaffolding is modeled at a more tractable level. The resulting reduced dimension Hessian matrix is advantageous as computer memory is limited. However, such a multiscale model is only useful if it can accurately describe both global (i.e., "low" frequency) and local (i.e., "high" frequency) motion at their respective resolutions.

3.4.1. Analyzing Normal Modes with SHAPES. The elastic network model (ENM) has proven able to qualitatively capture low-frequency, large-scale vibrational motion of protein systems.^{68,69} It is therefore desirable to use MSCALE to describe the bulk of a system using coarse grain methods (i.e., ENM), while important areas are treated atomistically.^{23,48,70} However, direct comparison between

the normal modes of different representations of a structure are problematic because the Hessians and normal mode vectors are different dimensions. Thus, it is not particularly easy to validate this method.

In order to overcome this difficulty, an analysis method was developed that is independent of the length of the mode vectors. This scheme makes use of the shape descriptor facility in CHARMM. Shape descriptors provide a convenient way to calculate the Cartesian moments of a given structure. For example, the three first-order moments are the mass-weighted averages of the x , y , and z coordinates (the coordinates can be weighted in many ways, but for this work, mass weighting was used exclusively). These values provide the center of mass of the system. Likewise, the six second-order moments are $\langle x^2 \rangle - \langle x \rangle^2$, $\langle y^2 \rangle - \langle y \rangle^2$, $\langle z^2 \rangle - \langle z \rangle^2$, $\langle xy \rangle - \langle x \rangle \langle y \rangle$, $\langle xz \rangle - \langle x \rangle \langle z \rangle$, and $\langle yz \rangle - \langle y \rangle \langle z \rangle$, which are the moments of inertia of the system.

For this analysis, coarse grained centers were treated as homogeneous spheres of a given radius. Changing the radius of any atom will only affect moments which contain solely even-order terms because the spatial extent of the spheres will increase or decrease by the same amount in the positive and negative directions. Therefore, the changes in the odd order terms will cancel each other out. For this work, a sphere radius of 4 Å was employed, as this value best reproduced the spatial extents of the all-atom systems. This is intuitive because in the ENM, beads are centered at C_α positions, and adjacent α carbons are generally slightly less than 4 Å away. It is therefore reasonable that 4 Å spheres were found to most closely represent the actual spatial extent of the all-atom system, including side chains.

For each of the modes being analyzed, the derivatives of the shape moments were calculated via finite differences. Large derivatives for low-order shape moments indicate large-scale global deformation of the system. For example, bending a helical structure aligned parallel to the z -axis in the x direction will result in a significant $\langle xz^2 \rangle$ moment change. To make a finite difference estimation, the starting structure was deformed by 0.01 Å along the mode vector, and the shape moments were generated for both the original and deformed structure up to the third order. The difference between the original and deformed moment provides an idea of how much the shape moments change for a very small movement along the normal mode. The first-order moments (and their differences) will be 0 if there is no net translation in the modes being studied.

In order to determine ideal weighting of the third-order moments relative to the second-order ones, the all-atom and ENM modes were generated for a test system, which consisted of a 31 residue α -helix that is described in more detail below. For each of these two representations, the dot products of the shape differences of the five lowest nonrotational/translational modes were calculated. This

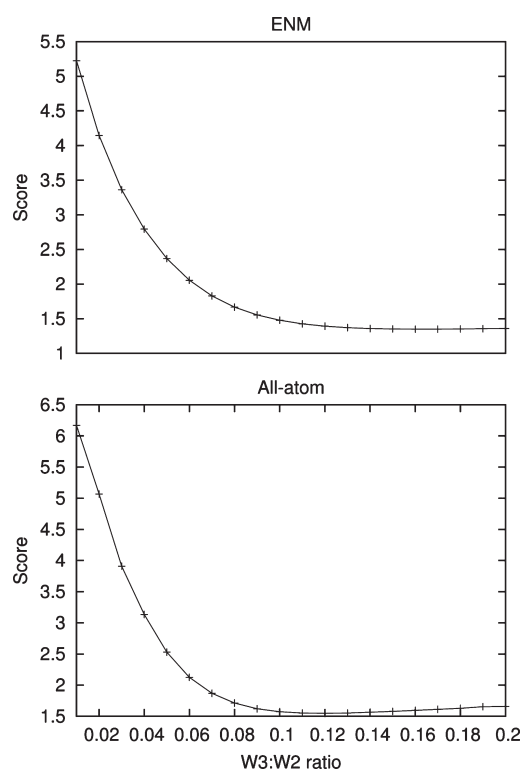


Figure 5. The sum of squares of the off-diagonal upper triangular elements of the 5×5 matrix obtained by dotting the normalized shape difference vectors against one another for the ENM and all-atom cases (see Section 3.4.1 for details) as a function of the weighting between the third- and second-order moments. Since off-diagonal elements are expected to be minimal, the optimal weighting was determined to be approximately 0.12 in each case but slightly higher for the ENM than the all-atom model.

means that the moments of all-atom mode 1 were treated as a vector, normalized, and dotted against the moments of modes 2–5. This was done for each of the five “shape difference vectors” corresponding to the motion of the modes, producing a symmetric 5×5 matrix. Since the shape differences do not form a mutually orthogonal basis set as the normal modes themselves do, the off-diagonal elements will not all be 0 (all diagonal elements will be 1).

A scoring function can be developed by taking the sum of squares of the off-diagonal upper triangle elements of the matrix. In the ideal case this scoring function would yield 0, indicating that the shape differences perfectly describe orthogonal motion of the structure. The value of this scoring function was calculated for various possible weightings of the third-order shape moments. The plot for a portion of this parameter space showing the minima for the all-atom and ENM cases is shown in Figure 5. The optimal parameter for this test system is that the third-order moments should be weighted by 0.135 relative to the second-order moments (the scoring function was optimal in the all-atom case when the weight was 0.12 and in the ENM case when it was 0.15, so the average of these two values was chosen).

3.4.2. Results. Normal mode analysis was performed on a 31 residue α -helix taken from the leucine zipper (PDB code 1GCL).⁷¹ This structure was minimized, and all-atom and ENM normal modes were generated. A force constant of 60 kcal/mol \AA^2 was used for adjacent centers, and a force constant of 6 kcal/mol \AA^2 was used for all other pairs of centers within a 10 \AA cutoff. These values were chosen so that the lowest modes yielded similar vibrational

Table 4. Dot Products of the Normalized Vectors Built from the Shape Moments of the Five Lowest Nonrotational/Translational Modes of the Given Representations of a 31 Residue α -Helix

	1	2	3	4	5
All-Atom vs ENM					
1	-0.797	0.583	-0.739	-0.217	-0.538
2	-0.689	-0.642	0.179	0.045	-0.232
3	0.027	0.720	-0.464	0.649	0.491
4	-0.417	-0.379	0.012	-0.747	-0.731
5	-0.300	-0.209	0.290	0.876	0.230
All-Atom vs All-Atom 4–31					
1	0.217	-0.970	0.230	-0.375	-0.156
2	-0.900	-0.112	-0.774	-0.709	0.022
3	0.208	-0.402	0.725	0.725	0.515
4	-0.058	-0.106	-0.506	-0.890	-0.573
5	-0.590	-0.098	-0.242	-0.069	0.918
All-Atom vs All-Atom 18–31					
1	-0.235	-0.937	0.849	0.216	-0.211
2	0.851	-0.394	-0.035	-0.556	-0.131
3	-0.614	-0.066	0.352	0.799	0.629
4	0.458	-0.402	0.124	-0.651	-0.714
5	0.454	-0.199	-0.323	0.438	0.844
All-Atom vs All-Atom 29–31					
1	-0.730	0.667	-0.715	-0.301	-0.731
2	-0.765	-0.567	0.318	0.046	-0.222
3	0.118	0.701	-0.614	0.619	-0.052
4	-0.470	-0.319	0.178	-0.763	-0.297
5	-0.299	-0.167	0.212	0.857	-0.060
ENM vs All-Atom 4–31					
1	0.407	0.784	0.295	0.675	-0.009
2	0.731	-0.651	0.878	0.500	0.004
3	-0.424	0.769	-0.657	-0.121	0.285
4	-0.375	0.009	0.089	0.411	0.938
5	-0.110	0.420	0.296	0.692	0.468
ENM vs All-Atom 18–31					
1	-0.351	0.928	-0.557	0.096	0.125
2	-0.900	-0.320	0.706	0.695	0.115
3	0.589	0.529	-0.958	-0.283	0.230
4	0.088	0.088	-0.314	0.643	0.962
5	-0.180	0.603	-0.295	0.259	0.526
ENM vs All-Atom 29–31					
1	0.993	-0.156	0.314	0.089	0.679
2	0.070	0.993	-0.875	-0.023	-0.296
3	0.323	-0.785	0.973	0.274	0.443
4	0.052	0.033	0.074	0.996	0.110
5	0.509	-0.045	0.073	0.568	0.820

frequencies to those of the all-atom structure. Next, three multiscale models of the structure were generated with: one model having residues 29–31 represented at the all-atom level (multiscale structure I), the next with residues 18–31 so treated (multiscale structure II), and the final one with residues 4–31 treated atomistically (multiscale structure III). The force constants for the ENM parts of these models were the same as those used to generate the pure ENM modes. The five lowest nonrotational/translational modes of each of the five mode sets (the all-atom, ENM, and three multiscale mode sets) were then generated and evaluated using the method described above. Shape derivatives for each of the 25 modes being studied were estimated via finite differences. The dot products of the shape derivatives for each of the lowest five modes of the three multiscale mode sets were then taken against the shape derivatives of the five lowest modes of both the all-atom and ENM mode sets. These are given in Table 4.

Table 5. Score Function of Off-Diagonal Elements for Each of the Three Multiscale Models Compared to the All-Atom and ENM Shapes^a

	all-atom	ENM
all-atom 29–31	3.440	3.053
all-atom 18–31	3.602	3.305
all-atom 4–31	3.394	3.436

^aThe score of the all-atom versus ENM shape derivatives is 3.722.

The results indicate that multiscale modes qualitatively reproduce the same types of deformations as the all-atom and ENM modes. When visualizing the motion of the structure as it is deformed along each mode, the five low-frequency modes represent either a bending or twisting movement of the helix. The modes which yield bending and twisting movements are not the same for each mode set, however, and this can be observed from the results. For example, looking at the comparison between the shape derivatives of the ENM structure and the multiscale structure II, the highest overlaps are reversed between modes one and two. Taking such reordering of modes into account, the shape derivatives for the multiscale structure where more residues are simulated at the all-atom level appear to overlap better with the shape derivatives for the all-atom structure; likewise, the shape derivatives for the structure where the fewest residues are simulated atomistically seem to overlap better with the ENM shape derivatives. The shape derivatives for multiscale structure II, which is split roughly half and half between all-atom and coarse grained representation, represent a middle ground.

In order to characterize these overlaps in a quantitative manner, a new scoring method was developed. For each row of the matrices given in Table 4, a score was generated by summing the squares of all of the elements except for the highest element in each row. This implicitly assumes that this element should be on the diagonal if there was no mode reordering. This method therefore yields the sum of squares of what would be the off-diagonal elements in this case. The results of the scoring function are given in Table 5. As expected, the shape derivatives of multiscale structure I have the best score function when compared to the ENM shape derivatives. Likewise, the shape derivatives of multiscale structure III score best against the all-atom shape derivatives. Interestingly, multiscale structure II is the worst scorer against the all-atom structure but scores in the middle against the ENM structure, as expected. This indicates that the low-frequency modes of this structure are much closer to those of the ENM structure than they are to the all-atom motions. As is to be expected, the scoring function is the worst when the ENM shape derivatives are compared to the all-atom ones. Furthermore, this method will not take into account motion that does not change the spatial extents of the molecule (e.g., a pure twisting motion of the helix that does not incorporate any bending).

4. CONCLUSIONS

This work introduces a general concurrent multiscale modeling framework and describes its implementation into CHARMM and several other classical and quantum mechanical packages. This approach, dubbed MSCALE, generalizes the ideas underlying both additive and subtractive multiscale schemes. Using a completely generalized implementation, this functionality allows

coupling of multiple, independent but connected calculations with each being a separate single or group of processes. This allows for arbitrary combinations of additive and subtractive schemes of any level of complexity. General symmetry and triclinicity is supported allowing periodic systems with constant pressure and temperature; an example will be given in future work. One of the most attractive features of the MSCALE framework is the relative ease of implementation, since MSCALE is based on the standard MPI communication protocol. This allows easy external access to the MSCALE framework, making the method widely available to the computational community. As multiscale modeling has increasingly become integral to biophysical simulations, the need for generalized and open software has likewise gained importance; MSCALE was created to fill this void and bring together functionality previously isolated in separate programs.

We report four examples that demonstrate the efficacy of the MSCALE approach and implementation. Perhaps the most straightforward use of such a tool is to perform subtractive QM/MM and QM/QM calculations (i.e., ONIOM-type). Although this is an easy way to apply multiscale modeling techniques, it is clear the limit of accuracy is at the low level of theory, and thus care must be taken to choose computational methods appropriately. Second, we detail the implementation of MSCALE with CHARMM's free energy perturbation module (PERT) and showcase the application of this with multiple force fields (CHARMM and AMBER99SB). Using this feature, vacuum and solvation free energies of the alanine dipeptide were computed. These were previously calculated for an earlier version of the AMBER force field in a previous study,⁶¹ however, MSCALE's functionality allowed the calculation to be performed in a more straightforward manner, using existing CHARMM functionality. Third, porting of the MSCALE communication paradigm was demonstrated by connecting AMBER's SANDER module to CHARMM and by using the implicit GB^{OBC} solvation model (only implemented in SANDER) with the CHARMM potential. Furthermore, an implementation of MSCALE for the TINKER program was developed, allowing CHARMM access to the AMOEBA force field through this interface. Finally, multiscale normal mode analysis (NMA) was carried out combining ENM and classical all-atom methodologies. This is interesting, as the algorithms needed to perform atomistic NMA have not kept pace with computational hardware. Although various techniques have been developed for simplifying the Hessian calculation of large systems, they generally require either numerical estimation or fixing or integrating out part of the system's motion. Therefore, combining unrestrained coarse grain and all-atom methods and achieving near atomistic quality results are highly desirable; this has been accomplished and demonstrated within the current framework.

Although this work significantly improves the tools available to the computational community, there is still a wide variety of possibilities for future work. Ongoing is an effort to integrate MSCALE with CHARMM's distributed replica methods (REPD), which will allow this facility to be used with chain-of-states or string methods within CHARMM^{72–74} or to facilitate the use of the MSCALE command with replica exchange. To give another example, there is much work that needs to be done to characterize the relationships between normal mode vectors produced by different multiscale models. As mentioned above, many analysis techniques do not deal with differently sized mode vectors, making these difficult to compare directly. Similar issues

arise for different types of calculations. Further consideration of the inherent trade-offs involved in multiscale modeling is therefore an important area of future study.

■ ASSOCIATED CONTENT

S Supporting Information. Example CHARMM input scripts showing how to set up calculations with the MSCALE command are provided. Additionally, a more technically oriented description of how the MSCALE implementation in CHARMM works is given, along with brief notes about making other codes MSCALE compatible. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: hlw@mail.usf.edu (H.L.W.); brb@nih.gov (B.R.B.)

Author Contributions

[†]These authors contributed equally.

■ ACKNOWLEDGMENT

H.L.W. would like to acknowledge NIH (1K22HL088341-01A1) and the University of South Florida (start-up) for funding. J.W.P. acknowledges funding from NIH (R01GM58712) and NSF (Cyberinfrastructure grant 0535675). This research was supported in part by the Intramural Research Program of the NIH, National Heart Lung and Blood Institute. The NHLBI funding for the LoBoS (<http://www.lobos.nih.gov>) cluster computing system is also acknowledged and appreciated.

■ REFERENCES

- Russel, D.; Lasker, K.; Phillips, J.; Schneidman-Duhovny, D.; Velzquez-Muriel, J. A.; Sali, A. *Curr. Opin. Cell Biol.* **2009**, *21*, 97–108.
- Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227–249.
- Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1986**, *7*, 718–730.
- Field, M. J.; Bash, P. A.; Karplus, M. *J. Comput. Chem.* **1990**, *11*, 700–733.
- Svensson, M.; Humbel, S.; Froese, R. D. J.; Matsubara, T.; Sieber, S.; Morokuma, K. *J. Phys. Chem.* **1996**, *100*, 19357–19363.
- Tozzini, V. *Acc. Chem. Res.* **2010**, *43*, 220–230.
- Sherwood, P.; Brooks, B. R.; Sansom, M. S. P. *Curr. Opin. Struct. Biol.* **2008**, *18*, 630–640.
- Miller, B. T.; Zheng, W.; Venable, R. M.; Pastor, R. W.; Brooks, B. R. *J. Phys. Chem. B* **2008**, *112*, 6274–6281.
- Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S. J. *Chem. Theory Comput.* **2008**, *4*, 819–834.
- Yap, E.; Fawzi, N. L.; Head-Gordon, T. *Proteins: Struct., Funct., Bioinf.* **2008**, *70*, 626–638.
- Yang, L.; Chng, C. *Bioinf. Biol. Insights* **2008**, *2*, 25–45.
- Chu, J.; Voth, G. A. *Biophys. J.* **2007**, *93*, 3860–3871.
- Zheng, W.; Brooks, B. R.; Thirumalai, D. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 7664–7669.
- Jeong, J. I.; Jang, Y.; Kim, M. K. *J. Mol. Graph. Mod.* **2006**, *24*, 296–306.
- Das, P.; Matysiak, S.; Clementi, C. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 10141–10146.
- Zheng, W.; Brooks, B. R. *Biophys. J.* **2005**, *88*, 3109–3117.
- Liwo, A.; Khalili, M.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 2362–2367.
- Cheung, M. S.; Garca, A. E.; Onuchic, J. N. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 685–690.

- Klimov, D. K.; Thirumalai, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 2544–2549.
- Gopal, S. M.; Mukherjee, S.; Cheng, Y.; Feig, M. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1266–1281.
- Maragakis, P.; Karplus, M. *J. Mol. Biol.* **2005**, *352*, 807–822.
- Tirion, M. M. *Phys. Rev. Lett.* **1996**, *77*, 1905–1908.
- Jernigan, R. L.; Bahar, I. *Curr. Opin. Struct. Biol.* **1996**, *6*, 195–209.
- Chennubhotla, C.; Rader, A. J.; Yang, L.; Bahar, I. *Phys. Biol.* **2005**, *2*, 173–180.
- Noid, W. G.; Chu, J.; Ayton, G. S.; Krishna, V.; Izvekov, S.; Voth, G. A.; Das, A.; Andersen, H. C. *J. Chem. Phys.* **2008**, *128*, 244114.
- Noid, W. G.; Liu, P.; Wang, Y.; Chu, J.; Ayton, G. S.; Izvekov, S.; Andersen, H. C.; Voth, G. A. *J. Chem. Phys.* **2008**, *128*, 244115.
- Chu, J.; Ayton, G. S.; Izvekov, S.; Voth, G. A. *Mol. Phys.* **2007**, *105*, 167–175.
- Vreven, T.; Byun, K. S.; Komromi, I.; Dapprich, S.; Montgomery, J. A.; Morokuma, K.; Frisch, M. J. *J. Chem. Theory Comput.* **2006**, *2*, 815–826.
- Senn, H. M.; Thiel, W. *Ang. Chem. Int. Ed.* **2009**, *48*, 1198–1229.
- Atomistic Approaches in Modern Biology*; Reiher, M., Ed.; Springer: Berlin and Heidelberg, Germany, 2007; Vol. 268.
- Brooks, B. R.; Brooks, C. L., III; Mackerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kucsera, 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.
- Geist, A.; Gropp, W.; Huss-Lederman, S.; Lumsdaine, A.; Lusk, E. L.; Saphir, W.; Skjellum, A.; Snir, M. MPI-2: Extending the Message-Passing Interface. In *Proceedings of Euro-Par*; 1996; Vol. I'96, pp 128–135.
- Ren, P.; Ponder, J. W. *J. Comput. Chem.* **2002**, *23*, 1497–1506.
- Ren, P.; Ponder, J. W. *J. Phys. Chem. B* **2003**, *107*, 5933–5947.
- 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.
- Valiev, M.; Bylaska, E.; Govind, N.; Kowalski, K.; Straatsma, T.; Dam, H. V.; Wang, D.; Nieplocha, J.; Apra, E.; Windus, T.; de Jong, W. *Comput. Phys. Commun.* **2010**, *181*, 1477–1489.
- 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.
- Crawford, T. D.; Sherrill, C. D.; Valeev, E. F.; Fermann, J. T.; King, R. A.; Leininger, M. L.; Brown, S. T.; Janssen, C. L.; Seidl, E. T.; Kenny, J. P.; Allen, W. D. *J. Comput. Chem.* **2007**, *28*, 1610–1616.
- 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.1; Gaussian, Inc.: Wallingford, CT, 2009.
- Shao, Y.; Molnar, L. F.; Jung, Y.; Kussmann, J.; Ochsenfeld, C.; Brown, S. T.; Gilbert, A. T.; Slipchenko, L. V.; Levchenko, S. V.; O'Neill,

- D. P.; R. A., D., Jr; Lochan, R. C.; Wang, T.; Beran, G. J.; Besley, N. A.; Herbert, J. M.; Lin, C. Y.; Voorhis, T. V.; Chien, S. H.; Sodt, A.; Steele, R. P.; Rassolov, V. A.; Maslen, P. E.; Korambath, P. P.; Adamson, R. D.; Austin, B.; Baker, J.; Byrd, E. F. C.; Dachsel, H.; Doerksen, R. J.; Dreuw, A.; Dunietz, B. D.; Dutoi, A. D.; Furlani, T. R.; Gwaltney, S. R.; Heyden, A.; Hirata, S.; Hsu, C.; Kedziora, G.; Khalliulin, R. Z.; Klunzinger, P.; Lee, A. M.; Lee, M. S.; Liang, W.; Lotan, I.; Nair, N.; Peters, B.; Proynov, E. I.; Pieniazek, P. A.; Rhee, Y. M.; Ritchie, J.; Rosta, E.; Sherrill, C. D.; Simmonett, A. C.; Subotnik, J. E.; Woodcock, H. L.; Zhang, W.; Bell, A. T.; Chakraborty, A. K.; Chipman, D. M.; Keil, F. J.; Warshel, A.; Hehre, W. J.; Schaefer, H. F.; Kong, J.; Krylov, A. I.; Gill, P. M. W.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3172–3191.
- (41) Schmidt, M. W.; Baldrige, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Kosecki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S. J.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347–1363.
- (42) Guest, M. F.; Bush, I. J.; van Dam, H. J. J.; Sherwood, P.; Thomas, J. H. H.; van Lenthe, J. H.; Havenith, R. W. A.; Lendrick, J. *Mol. Phys.* **2005**, *103*, 719–747.
- (43) Cui, Q.; Elstner, M.; Kaxiras, E.; Frauenheim, T.; Karplus, M. *J. Phys. Chem. B* **2001**, *105*, 569–585.
- (44) Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4899–4907.
- (45) Walker, R. C.; Crowley, M. F.; Case, D. A. *J. Comput. Chem.* **2008**, *29*, 1019–1031.
- (46) Becker, D.; Ligon, W.; Merkey, P.; Ross, R. *IEEE Software* **1999**, *16*, 79.
- (47) Dapprich, S.; Komaromi, I.; Byun, K.; Morokuma, K.; Frisch, M. J. *J. Mol. Struct. (THEOCHEM)* **1999**, *461*, 1–21.
- (48) Ghysels, A.; Woodcock, H. L.; Larkin, J. D.; Miller, B. T.; Shao, Y.; Kong, J.; van Neck, D.; van Speybroek, V.; Waroquier, M.; Brooks, B. R. *J. Chem. Theory Comput.* **2011**, *7* (2), 496–514.
- (49) Straatsma, T. P.; Berendsen, H. J. C. *J. Chem. Phys.* **1988**, *89*, 5876.
- (50) Woodcock, H. L.; Hodoscek, M.; Gilbert, A. T. B.; Gill, P. M. W.; Schaefer, H. F.; Brooks, B. R. *J. Comput. Chem.* **2007**, *28*, 1485–1502.
- (51) MacKerell, A. D.; Brooks, B.; Brooks, C. L.; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M. In *Encyclopedia of Computational Chemistry*; v. R. Schleyer, P., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., Schreiner, P. R., Eds.; John Wiley & Sons, Ltd: Chichester, U.K., 2002.
- (52) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 712–725.
- (53) Wei, Y.; Latour, R. A. *Langmuir* **2009**, *25*, 5637–5646.
- (54) Fears, K. P.; Sivaraman, B.; Powell, G. L.; Wu, Y.; Latour, R. A. *Langmuir* **2009**, *25*, 9319–9327.
- (55) Sivaraman, B.; Latour, R. A. *Biomaterials* **2010**, *31*, 832–839.
- (56) Vellore, N. A.; Yancey, J. A.; Collier, G.; Latour, R. A.; Stuart, S. J. *Langmuir* **2010**, *26*, 7396–7404.
- (57) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press: Oxford, England, 1987.
- (58) Louwerse, M. J.; Baerends, E. J. *Chem. Phys. Lett.* **2006**, *421*, 138.
- (59) Klauda, J. B.; Pastor, R. W.; Brooks, B. R. *J. Phys. Chem. B* **2005**, *109*, 15684–15686.
- (60) Das, D.; Eurenus, K. P.; Billings, E. M.; Sherwood, P.; Chatfield, D. C.; Hodoscek, M.; Brooks, B. R. *J. Chem. Phys.* **2002**, *117*, 10534.
- (61) König, G.; Bruckner, S.; Boresch, S. *J. Comput. Chem.* **2009**, *30*, 1712–1718.
- (62) Bennett, C. H. *J. Comput. Phys.* **1976**, *22*, 245–268.
- (63) Ponder, J. W.; Richards, F. M. *J. Comput. Chem.* **1987**, *8*, 1016–1024.
- (64) Kundrot, K. E.; Ponder, J. W.; Richards, F. M. *J. Comput. Chem.* **1991**, *12*, 402–409.
- (65) Onufriev, A.; Bashford, D.; Case, D. A. *Proteins: Struct., Funct., Bioinf.* **2004**, *55*, 383–394.
- (66) Ryckaert, J.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.
- (67) Andersen, H. C. *J. Comput. Phys.* **1983**, *52*, 24–34.
- (68) Bahar, I.; Rader, A. *Curr. Opin. Struct. Biol.* **2005**, *15*, 586–592.
- (69) Haliloglu, T.; Bahar, I.; Erman, B. *Phys. Rev. Lett.* **1997**, *79*, 3090–3093.
- (70) Woodcock, H. L.; Zheng, W.; Ghysels, A.; Shao, Y.; Kong, J.; Brooks, B. R. *J. Chem. Phys.* **2008**, *129*, 214109.
- (71) Harbury, P. B.; Zhang, T.; Kim, P. S.; Alber, T. *Science* **1993**, *262*, 1401–1407.
- (72) Woodcock, H. L.; Hodoscek, M.; Sherwood, P.; Lee, Y. S.; Schaefer, H. F.; Brooks, B. R. *Theor. Chem. Acc.* **2003**, *109*, 140–148.
- (73) Chu, J. W.; Trout, B. L.; Brooks, B. R. *J. Chem. Phys.* **2003**, *119*, 12708.
- (74) Woodcock, H. L.; Hodoscek, M.; Brooks, B. R. *J. Phys. Chem. A* **2007**, *111*, 5720–5728.

Optimizing Protein–Solvent Force Fields to Reproduce Intrinsic Conformational Preferences of Model Peptides

Paul S. Nerenberg[†] and Teresa Head-Gordon^{*,†,‡}

[†]California Institute of Quantitative Biosciences (QB3), [‡]Department of Bioengineering, University of California, Berkeley, Berkeley, California 94720-3220, United States

S Supporting Information

ABSTRACT: While most force field efforts in biomolecular simulation have focused on the parametrization of the protein, relatively little attention has been paid to the quality of the accompanying solvent model. These considerations are especially relevant for simulations of intrinsically disordered peptides and proteins, for which energy differences between conformations are small and interactions with water are enhanced. In this work, we investigate the accuracy of the AMBER ff99SB force field when combined with the standard TIP3P model or the more recent TIP4P-Ew water model, to generate conformational ensembles for disordered trialanine (Ala₃), triglycine (Gly₃), and trivaline (Val₃) peptides. We find that the TIP4P-Ew water model yields significantly better agreement with experimentally measured scalar couplings—and therefore more accurate conformational ensembles—for both Ala₃ and Gly₃. For Val₃, however, we find that the TIP3P and TIP4P-Ew ensembles are equivalent in their performance. To further improve the protein–water force field combination and obtain more accurate intrinsic conformational preferences, we derive a straightforward perturbation to the ϕ' backbone dihedral potential that shifts the β –PPII equilibrium. We find that the revised ϕ' backbone dihedral potential yields improved conformational ensembles for a variety of small peptides and maintains the stability of the globular ubiquitin protein in TIP4P-Ew water.

INTRODUCTION

Over the past three decades, classical molecular dynamics (MD) and Monte Carlo (MC) simulations have emerged as an important complement to experimental methods for investigating many aspects of biomolecular structure, dynamics, and function.¹ The predictive quality of biomolecular simulation depends on the accuracy of the potential energy function (or force field), comprised of bonded interactions, van der Waals interactions, and (typically) fixed-charge electrostatics, many of which were parameterized approximately 15 years ago.^{2–4} Bonded parameters describing bond stretching, bending, and torsions and partial atomic charges were generally derived using quantum chemical methods, and the van der Waals parameters were empirically derived to match experimental densities and enthalpies of vaporization of neat organic liquids (e.g., liquid hydrocarbons) in an effort to model the intramolecular interactions typical of folded globular proteins.^{2–4} Since that time, force field development efforts have utilized more advanced *ab initio* methods and focused on improving the backbone and side chain dihedral angle potentials^{5–10} and the fixed charges^{11,12} by minimizing the differences between gas phase *ab initio* and molecular mechanics energies, charge distributions, etc., for specified molecular structures. The ongoing refinement of empirical biomolecular force fields has resulted in a corresponding increase in the predictive power of MD simulations¹³ in modeling the structures, dynamics, and folding of globular proteins.^{14–16}

In the past 10 years, however, there has been a growing recognition that much of the human proteome (~30%) consists of proteins that are intrinsically disordered.^{17–19} Intrinsically disordered proteins (IDPs) are thought to be vital for carrying out a variety of signaling and regulatory functions in the cell, but

they have also been implicated in several common diseases, including various cancers, Alzheimer's disease, Parkinson's disease, type II diabetes, and cardiovascular disease.^{17,18,20} As it is known that IDPs rapidly sample multiple conformations (i.e., faster than the millisecond time scale of NMR experiments), the free energy landscapes of these proteins must be relatively flat, with small energetic barriers between conformations and energy differences that are on the order of $k_B T$.²¹ This differs markedly from the single, deep free energy minimum (folding funnel) that is characteristic of folded proteins^{22,23} and presents a potentially serious challenge to the predictive power of MD force fields when applied to such systems. In addition, by virtue of having greater net charge per residue and proportionally fewer hydrophobic residues, IDPs are generally more unfolded and solvent-exposed than their globular counterparts.^{19,24} Because of their increased solvent exposure, the conformational ensembles of IDPs likely depend more strongly on sensitive protein–water interactions, and therefore deficiencies of the standard TIP3P or SPC solvent models may become more obvious in simulations of these systems.²⁵

Recent studies enabled by improvements in both simulation methodologies and computer hardware and the development of experimental methods that yield high quality quantitative data (e.g., NMR scalar couplings) have enabled direct comparisons between simulation results and experimental results. These comparisons have revealed discrepancies in the backbone and side chain conformational preferences of short peptides^{8,26,27} and solvation free energies of small molecules,^{12,28,29} lending credence

Received: January 7, 2011

Published: March 07, 2011

to the notion that current force fields may not be optimal for simulations of more solvent-exposed peptides and IDPs. Related to the work presented here, Best and Hummer analyzed a number of different protein force fields and concluded that most are too helical when comparing calculated J-coupling observables to experiment, although the AMBER ff99SB⁷ force field proved to be more reliable than most.²⁶ In subsequent work, they developed a correction to the ψ backbone dihedral angle potential for AMBER ff99SB, deemed ff99SB*, which improved the helix-to-coil transition for longer α -helix-forming peptides.⁵ However, the only water model considered in that study was the TIP3P model,³⁰ which continues to dominate the biomolecular simulation literature in spite of demonstrable improvements in condensed phase water descriptions by other (potentially) compatible fixed charge water models.^{31–34}

Evidence that improved water models can yield more accurate conformational ensembles—particularly of disordered peptides—has been mounting in the past few years. A study of the A β _{21–30} peptide by Fawzi et al. found that the combination of AMBER ff99SB and the recently developed TIP4P-Ew model³³ yielded better predictions of NMR ROESY crosspeaks than ff99SB and TIP3P.³⁵ Using the same force field, Wickstrom et al. demonstrated that ensembles generated with TIP4P-Ew predicted NMR scalar couplings for Ala₃ and Ala₅ more accurately than ensembles generated with TIP3P.²⁷ More recently, Best and Mittal used a methodology similar to the ff99SB* study to develop a correction to the ψ backbone dihedral angle potential for AMBER ff03¹¹ with the TIP4P/2005 water model.³¹ They found that the optimized force field and solvent model combination (ff03w) yielded a more cooperative helix-to-coil transition and a more realistic collapse of unfolded conformers with increasing temperature than an optimized combination of ff03 and TIP3P water (ff03*).⁶

Therefore, we set out to understand how well a modern force field, AMBER ff99SB, and two water models, the default TIP3P model and the newer TIP4P-Ew model, could quantify the conformational ensembles of three disordered tripeptides (Ala₃, Gly₃, and Val₃) for which there are a large number of experimentally measured NMR scalar couplings at 300 K.³⁶ Additionally, we simulate the Ala₃ system at 275, 325, and 350 K—temperatures for which the same coupling data are available³⁶—to evaluate how the temperature-dependent characteristics of the water model (and protein force field) impact the simulated conformational ensembles.

Our results suggest that the use of the TIP4P-Ew water model, which is known to reproduce several characteristics of water better than TIP3P,³³ produces superior conformational ensembles for both the Ala₃ and Gly₃ peptides; for the Val₃ peptide, the difference in accuracy between the TIP3P and TIP4P-Ew ensembles is less significant. Because there are few cooperative interactions, e.g., hydrogen bonds and dipole alignments to stabilize secondary structure in these short peptides, we hypothesize that the backbone dihedral potentials likely play the primary role in determining the conformational ensembles. We therefore explore whether optimizing the backbone dihedral potentials to reproduce intrinsic conformational preferences of single amino acids—again in TIP4P-Ew water—can improve the conformational ensembles. By optimizing a single term of the ϕ' potential of the ff99SB force field with respect to NMR scalar couplings measured on a series of GXG peptides,³⁷ we find that we can broadly improve intrinsic conformational preferences in disordered peptides (GLG, Ala₃, and Val₃) without disrupting the

excellent native state stability of the globular protein ubiquitin. Moreover, these studies suggest several avenues for future improvements to the AMBER ff99SB force field, or fixed-charge force fields in general, when combined with more current water models such as TIP4P-Ew.

METHODS

Charges for Protonated C-Terminal Residues. All of the NMR scalar couplings referenced in this work were obtained at pH 2,^{36,37} which necessitates the use of protonated C-terminal residues (–COOH) in the MD simulations. Charges for C-terminal Ala and Val residues were derived by generating three conformations (corresponding to α , β , and PPII), optimizing these conformations at the HF/6-31G* level of theory, and then calculating the electrostatic potentials of these structures at the same level of *ab initio* theory.² All *ab initio* calculations were performed using GAMESS-US.³⁸ Charges were fit to these potentials using the RESP method,³⁹ as implemented in RED Server, version 1.0.⁴⁰ The charge derivation for the C-terminal Gly residue followed the same procedure except that five conformations were used instead of three. Charges for all three C-terminal residues are listed in the Supporting Information.

Simulation Protocol. Ala₃, Gly₃, Val₃, and GXG peptides were built in an extended conformation using the *tleap* program included with AmberTools 1.4.⁴¹ Each peptide was solvated in a truncated octahedron of 665–667 TIP3P³⁰ or TIP4P-Ew³³ water molecules, and the solvated system was neutralized by the addition of a single Cl[–] ion.⁴² All simulations were performed using AMBER 11 with either the ff99SB⁷ or ff99SB*⁵ force fields. Periodic boundary conditions were employed with a 9 Å cutoff for direct-space nonbonded interactions. Long-range electrostatics were calculated using particle mesh Ewald (PME)^{43,44} with default parameters for grid spacing and spline interpolation, and an analytic correction was employed for the van der Waals interactions beyond the cutoff. Dynamics were conducted with a 2 fs time step, and all bonds involving hydrogen atoms were constrained with SHAKE.⁴⁵ First, each system was minimized with 500 steps of steepest descent minimization, followed by 1000 steps of conjugate gradient minimization. The system was then equilibrated at 300 K for 50 ps using a Langevin thermostat with a coupling constant of 0.5 ps^{–1}. During both minimization and heating, peptide atoms were restrained with a force constant of 10.0 kcal mol^{–1} Å^{–2}. The system was then brought to appropriate density by equilibrating at a constant temperature and pressure of 300 K and 1.0 bar, respectively, for 250 ps. This NPT equilibration was performed using a Langevin thermostat coupling constant of 1.0 ps^{–1} and a Berendsen barostat coupling constant of 5.0 ps^{–1}.⁴⁶ Four independent starting configurations were generated by simulating the equilibrated system in the NVT ensemble at 400 K for 10 ns and drawing four conformations evenly from the last 8 ns. Each of these configurations was then equilibrated for 250 ps at 300 K (or 275, 325, or 350 K for Ala₃) and 1.0 bar. Four production simulations were then performed in the NPT ensemble for 100–400 ns, depending on the convergence properties of the simulation (discussed in greater detail in the Results section). Structures were saved every 1 ps. Definitions in terms of ϕ/ψ regions for the α , β , and PPII conformations are taken from Best et al.²⁶

Replica Exchange MD Simulations. To improve convergence of the simulation data for the low temperature (275 K) Ala₃ and Val₃ systems, two independent reservoir replica

exchange MD (RREMD) simulations^{47,48} were performed. For Ala₃, starting configurations were generated by first equilibrating the system at 275 K and then performing a 400 K NVT simulation, as described above. For both Ala₃ and Val₃, starting configurations were drawn from the first and fourth structures obtained by the 400 K simulations. Structure reservoirs of 50 000 structures were generated by simulating the system at 380 K (Ala₃) or 400 K (Val₃) for 50 ns and saving conformations every 1 ps. For Ala₃, 16 replicas of the system were equilibrated at exponentially spaced temperatures ranging from 275.00 to 372.40 K for 250 ps. For Val₃, 14 replicas of the system were equilibrated at exponentially spaced temperatures ranging from 300.0 to 391.9 K for 250 ps. RREMD simulations were then performed with the 380 K (Ala₃) or 400 K (Val₃) reservoirs for 50 ns, with swaps attempted between neighboring replicas every 1 ps. This temperature spacing yielded acceptance ratios of approximately 30–45%. An identical simulation protocol was used to simulate the Val₃ peptides with two different sets of modified van der Waals parameters, as well as with the modified backbone potential described in the Results section.

For GXG peptides, 20 replicas of the system were equilibrated in the NVT ensemble at exponentially spaced temperatures ranging from 298 to 450 K for 250 ps. REMD simulations⁴⁸ were then performed for 50 ns, with swaps attempted between neighboring replicas every 1 ps. The temperature spacing yielded exchange probabilities of approximately 28–40%. When optimizing the backbone parameters, this simulation protocol was carried out for four equally spaced values of the $n = 2$ ϕ' backbone dihedral angle potential ranging from a barrier height of 2.00 (ff99SB value) down to 1.55 kcal/mol. Piecewise cubic Hermite polynomials were used to interpolate the resulting data from 0.15 to 0.05 kcal/mol intervals.

To validate the optimized $n = 2$ ϕ' (i.e., C–N–C α –C β) backbone dihedral angle potential, we performed REMD simulations of the GLG and Ala5 peptides with both the unmodified and modified ff99SB force fields. The GLG simulations were carried out as described above, but using two different starting conformations: a fully extended conformation ($\phi = 180^\circ$, $\psi = 180^\circ$) and an α -helical conformation ($\phi = -60^\circ$, $\psi = -45^\circ$). The Ala₅ peptides were simulated with the same basic protocol but were instead solvated with 902 TIP4P-Ew water molecules and used 24 exponentially spaced replicas instead of 20 to account for the larger number of degrees of freedom. Exchange probabilities for the Ala₅ system ranged from 31 to 43%.

MD Simulations of Ubiquitin. A native state structure for ubiquitin was obtained from the PDB crystal structure 1UBQ.⁴⁹ Hydrogen atoms were added by tleap, and the sole histidine residue was protonated to be consistent with the NMR relaxation data, which were obtained at pH 4.7.⁵⁰ The system was solvated in a truncated octahedron of 3602 TIP4P-Ew water molecules and neutralized with 1 Na⁺ and 2 Cl[−] ions, consistent with the experimental salt concentration of 10 mM NaCl.⁵⁰ Simulations were performed with both the unmodified and modified ff99SB force fields. The solvated system was first minimized with 500 steps of steepest descent minimization, followed by 1500 steps of conjugate gradient minimization, using Cartesian restraints on the protein atoms with a force constant of 10.0 kcal mol^{−1} Å^{−2}. The entire system was then minimized again with the same number of steps, except without any restraints on the protein atoms. Next, the system was equilibrated in the NVT ensemble at 298 K for 50 ps using a Langevin thermostat with a coupling constant of 0.5 ps^{−1}. The system was then brought to appropriate

density by equilibrating at a constant temperature and pressure of 298 K and 1.0 bar, respectively, for 250 ps. During both NVT and NPT equilibration, the protein atoms were restrained with a force constant of 10.0 and 2.0 kcal mol^{−1} Å^{−2}, respectively. The system was then equilibrated in the NPT ensemble for an additional 5 ns without any restraints. Production simulations were run for 60 ns, with structures saved every 1 ps.

Generalized Order Parameters. Assuming that the slower overall motion of ubiquitin is isotropic and independent of faster internal motions,⁵¹ we eliminated rigid body rotations from the ubiquitin trajectories by performing a mass-weighted all-atom RMS fit using the first frames of the trajectories as reference structures. Next, we calculated the time autocorrelation function for the NH bond vectors:

$$C_I(t) = \langle P_2(\hat{\mu}(0) \cdot \hat{\mu}(t)) \rangle \quad (1)$$

where $P_2(x)$ is the second Legendre polynomial and $\hat{\mu}(t)$ is the unit vector of the NH bond. We then fit these correlation functions with the simplest approximation for internal motion:⁵¹

$$C_I(t) = S^2 + (1 - S^2) e^{-t/\tau_{\text{eff}}} \quad (2)$$

to determine the S^2 value for each bond vector.

RESULTS

A variety of NMR scalar couplings were calculated for Ala₃ using the ϕ/ψ backbone dihedral angles measured from the structures generated by the MD and RREMD simulations, and compared to the experimental coupling measurements of Graf et al.³⁶ that probe the ψ_1 , ϕ_2 , ψ_2 , and ϕ_3 dihedral angles. In particular, we calculated $^3J(\text{H}_N, \text{H}_\alpha)$, $^3J(\text{H}_N, \text{C})$, $^3J(\text{H}_\alpha, \text{C}')$, $^3J(\text{H}_N, \text{C}_\beta)$, $^3J(\text{H}_N, \text{C}_\alpha)$, $^1J(\text{N}, \text{C}_\alpha)$, and $^2J(\text{N}', \text{C}_\alpha)$ couplings, using three different sets of Karplus equation parameters (“orig.,” “DFT1,” and “DFT2”).²⁶ For Gly₃ and Val₃, we also calculated the $^3J(\text{C}, \text{C}')$ coupling, again using the same Karplus equation parameters as Best et al.²⁸

Similarly, we calculated the overall error between the calculated and experimental couplings as

$$\chi^2 = \frac{1}{N} \sum_{i=1}^N \frac{(\langle J_i \rangle_{\text{calc}} - J_{i, \text{expt}})^2}{\sigma_i^2} \quad (3)$$

As with previous studies, we assumed that errors in the calculated couplings due to sampling and errors in the experimentally measured couplings were negligible and therefore that the primary source of error (σ_i) for each coupling was the Karplus equation parameters themselves.^{5,26,27} The exact values of the coupling errors are given in the Supporting Information. We conservatively increased the error estimates in the Karplus equation parameters for the $^3J(\text{H}_N, \text{H}_\alpha)$, $^3J(\text{H}_N, \text{C})$, $^3J(\text{H}_\alpha, \text{C}')$, $^3J(\text{H}_N, \text{C}_\beta)$, and $^3J(\text{C}, \text{C}')$ couplings by 10% to account for the fact that the values given in the literature are mean absolute deviations, as opposed to root-mean-square deviations.²⁶ For comparison, previous studies increased these error estimates by 30% for the same reason.^{5,26,27} Thus, we expect our χ^2 values to be generally larger than those found in previous studies due to the use of generally smaller error estimates.

Simulations of Ala₃ at Multiple Temperatures. We carried out four independent MD simulations at 275, 300, 325, and 350 K and compared the results obtained with ff99SB and either the TIP3P or the TIP4P-Ew water model. In addition, for select temperatures, we also used the newly developed ff99SB* force

Table 1. χ^2 Values for Calculated Scalar Couplings of Ala₃ at (a) 350 K, (b) 325 K, (c) 300 K, and (d) 275 K for Various Force Field and Water Model Combinations^a

(a) 350 K			
	orig.	DFT1	DFT2
ff99SB (TIP3P)	1.82 (0.09)	1.30 (0.03)	1.40 (0.09)
ff99SB (TIP4P-Ew)	1.60 (0.06)	1.18 (0.01)	1.16 (0.06)
ff99SB* (TIP4P-Ew)	1.54 (0.05)	1.20 (0.02)	1.11 (0.05)
(b) 325 K			
	orig.	DFT1	DFT2
ff99SB (TIP3P)	1.63 (0.07)	0.97 (0.05)	1.21 (0.07)
ff99SB (TIP4P-Ew)	1.39 (0.04)	0.87 (0.04)	0.96 (0.04)
(c) 300 K			
	orig.	DFT1	DFT2
ff99SB (TIP3P)	2.38 (0.21)	1.21 (0.13)	1.80 (0.21)
ff99SB (TIP4P-Ew)	1.90 (0.20)	0.99 (0.09)	1.31 (0.19)
(d) 275 K			
	orig.	DFT1	DFT2
ff99SB (TIP3P)	2.92 (0.16)	1.97 (0.10)	2.51 (0.15)
ff99SB (TIP3P, RREMD)	2.76 (0.54)	1.91 (0.35)	2.36 (0.54)
ff99SB (TIP4P-Ew)	1.95 (0.33)	1.53 (0.13)	1.59 (0.29)
ff99SB (TIP4P-Ew, RREMD)	2.14 (0.07)	1.49 (0.06)	1.72 (0.08)
ff99SB* (TIP4P-Ew, RREMD)	2.71 (0.13)	1.90 (0.06)	2.30 (0.11)

^a Values are given as the means over four independent simulations, with the standard errors of the means given in parentheses. For RREMD simulations, probabilities are given as means over two independent simulations, with the differences between the two simulations given in parentheses.

field of Best and Hummer,⁵ together with the TIP4P-Ew water model.

At 350 K, our data demonstrate that both ff99SB and ff99SB*, together with the TIP4P-Ew water model, yield conformational ensembles more consistent with experimental data than ff99SB with the TIP3P water model (Table 1a). Moreover, this result is independent of the Karplus equation parameters used. A comparison between the ff99SB and ff99SB* results reveals that the use of a different water model brings about a larger change in the ensemble than the use of a modestly different force field (Table 1a). The primary difference between the TIP3P and TIP4P-Ew ensembles is an increase in the extended (β and PPII) conformations of the central Ala residue relative to more compact α -helical conformations (Table 2a). The ff99SB* force field is somewhat more helical than ff99SB, but the change in water model again imparts a larger difference than the change in force field (Table 2a). It is important to note, however, that the central residue's conformation does not account for couplings that measure the ψ_1 or ϕ_3 dihedral angles, which we examine separately below.

The TIP4P-Ew water model also generates more accurate ensembles at both 300 and 325 K (Table 1b and c), again showing a higher

propensity to sample β and PPII conformations than TIP3P. Interestingly, the β propensity remains relatively constant with decreasing temperature, while the α propensity decreases and the PPII propensity increases—regardless of the solvent model used (Table 2b and c).

The results at 275 K provide the most sensitive test of the force field—water model combinations, as relative differences in energies more strongly contribute to differences in ensembles (via their Boltzmann weights) at this low temperature. We again found that TIP4P-Ew yielded demonstrably better conformational ensembles at this temperature than TIP3P, with the difference being even more pronounced than at higher temperatures (Table 1d). As before, the primary difference between the TIP3P and TIP4P-Ew ensembles is an increase in the extended (primarily PPII) conformations of the central Ala residue relative to more compact α -helical conformations (Table 2d).

Because convergence at low temperatures is difficult, we also conducted two independent RREMD simulations to corroborate the results of the “conventional” MD simulations. Although the RREMD simulations are performed in the NVT ensemble, as opposed to the NPT ensemble employed in the conventional MD simulations, differences between the two are likely minimal at the target temperature (275 K). The results of the RREMD simulations again confirm that TIP4P-Ew yields better ensembles for Ala₃ than TIP3P (Table 1d). In addition, we note that the ff99SB/TIP4P-Ew combination is significantly more accurate than the ff99SB*/TIP4P-Ew combination at this low temperature (Table 1d). While there is little difference between these combinations in the conformational preferences of the central residue (Table 1d), the behavior of the N- and C-terminal residues differs significantly, with ff99SB* stabilizing α -helical conformations of the N-terminal residue (Supporting Information Figure 1a) and turn/ α_L conformations of the C-terminal residue (Supporting Information Figure 1b).

An examination of individual scalar couplings from the 275 K TIP3P and TIP4P-Ew data suggests that much of the observed improvement is due to a decrease in the sampling of the β conformation by the ϕ angle of the third residue and to a lesser extent the second residue, indicated by a decrease in magnitudes of the $^3J(H_{N_i}, H_{\alpha_i})$, $^3J(H_{N_i}, C_i)$, and $^3J(H_{\alpha_i}, C_i')$ couplings and an increase in the magnitude of the $^3J(H_{N_i}, C_{i\beta})$ coupling (Supporting Information Tables 1–3). There is also a decrease in sampling of the α conformation by the ψ angles of the first and second residue, which is indicated by an increase in magnitude of the $^3J(H_{N_i}, C_{i\alpha})$ and $^2J(N_i', C_{i\alpha})$ couplings (Supporting Information Tables 1–3). These observations are consistent with the central residue data that suggest an increase in the sampling of the PPII conformation in TIP4P-Ew water relative to TIP3P (Table 2d) and that the relative stabilization of the PPII conformation in TIP4P-Ew is primarily responsible for its improved performance relative to TIP3P.

Simulations of Gly₃ at 300 K. We performed four independent simulations of Gly₃ at 300K, using both TIP3P and TIP4P-Ew water, as well as the modified ff99SB* force field with TIP4P-Ew water. For all three force field and solvent model combinations, we observed large discrepancies between the observed and calculated scalar couplings, leading to χ^2 values of 2.93–3.45 (Table 3). Nonetheless, both of the TIP4P-Ew simulations had consistently lower χ^2 values than the TIP3P simulation, correlated with slightly greater sampling of the PPII conformation (Supporting Information Table 4), while there was little difference between ff99SB and ff99SB* (Table 3). These data suggest that the TIP4P-Ew water model again results in a more accurate conformational ensemble than TIP3P due to enhanced sampling of the PPII conformation.

Table 2. Conformational Preferences of the Central Residue of Ala₃ at (a) 350 K, (b) 325 K, (c) 300 K, and (d) 275 K for Various Force Field and Water Model Combinations^a

	(a) 350 K			
	α	β	PPII	other
#99SB (TIP3P)	0.169 (0.008)	0.390 (0.002)	0.392 (0.002)	0.048 (0.006)
#99SB (TIP4P-Ew)	0.118 (0.006)	0.420 (0.005)	0.426 (0.006)	0.036 (0.012)
#99SB* (TIP4P-Ew)	0.139 (0.007)	0.412 (0.006)	0.413 (0.004)	0.036 (0.012)
	(b) 325 K			
	α	β	PPII	other
#99SB (TIP3P)	0.142 (0.004)	0.391 (0.006)	0.422 (0.002)	0.044 (0.014)
#99SB (TIP4P-Ew)	0.100 (0.006)	0.416 (0.004)	0.456 (0.006)	0.028 (0.008)
	(c) 300 K			
	α	β	PPII	other
#99SB (TIP3P)	0.121 (0.004)	0.396 (0.004)	0.450 (0.002)	0.032 (0.005)
#99SB (TIP4P-Ew)	0.081 (0.002)	0.411 (0.004)	0.488 (0.005)	0.021 (0.006)
	(d) 275 K			
	α	β	PPII	other
#99SB (TIP3P)	0.102 (0.009)	0.390 (0.004)	0.476 (0.006)	0.032 (0.012)
#99SB (TIP3P, RREMD)	0.106 (0.004)	0.388 (0.017)	0.479 (0.007)	0.027 (0.020)
#99SB (TIP4P-Ew)	0.068 (0.005)	0.402 (0.002)	0.519 (0.005)	0.011 (0.007)
#99SB (TIP4P-Ew, RREMD)	0.069 (0.015)	0.399 (0.007)	0.511 (0.004)	0.021 (0.003)
#99SB* (TIP4P-Ew, RREMD)	0.063 (0.011)	0.402 (0.008)	0.526 (0.003)	0.009 (0.005)

^a Values are given as the means over four independent simulations, with the standard errors of the means given in parentheses. For RREMD simulations, probabilities are given as means over two independent simulations, with the differences between the two simulations given in parentheses.

Table 3. χ^2 Values for Calculated Scalar Couplings of Gly₃ at 300 K for Various Force Field and Water Model Combinations^a

	all couplings			no ² J(N',C _α) or ³ J(C,C') coupling		
	orig.	DFT1	DFT2	orig.	DFT1	DFT2
#99SB (TIP3P)	3.21 (0.03)	3.45 (0.05)	3.26 (0.04)	0.57 (0.02)	1.14 (0.04)	0.73 (0.03)
#99SB (TIP4P-Ew)	2.93 (0.04)	3.11 (0.08)	2.96 (0.05)	0.47 (0.05)	0.98 (0.11)	0.62 (0.06)
#99SB* (TIP4P-Ew)	2.92 (0.02)	3.08 (0.06)	2.93 (0.03)	0.46 (0.02)	0.96 (0.06)	0.60 (0.03)

^a Values are given as the means over four independent simulations, with the standard errors of the means given in parentheses.

Graf et al. have suggested that there may be errors in the Karplus equation parameters due to the limited number of measurements on glycine residues.³⁶ To better understand the cause of the large discrepancies in the scalar couplings, we first examined which of the calculated couplings were making the largest contribution to the χ^2 values. We found that two couplings (out of a total of 12)—the ²J(N',C_α) and ³J(C,C') couplings of the central residue—approximately equally contributed 60–80% of the total χ^2 value, with the variation due to the different sets of Karplus equation parameters used. In addition, the ²J(N',C_α) coupling of the third residue and the ³J(H_ω,C') coupling of the central residue made lesser contributions to the overall error.

The underlying cause of the discrepancies in the ²J(N',C_α) couplings appears to be due to a residue-specific effect that renders the Karplus equation parameters for this coupling inaccurate. More precisely, the maximum possible value of the coupling using these parameters is 8.71 Hz (with an uncertainty of ±0.5 Hz), while the experimental values are 10.45 and 9.05

Hz, respectively, for second and third residues of Gly₃. Thus, even if the simulation-generated ensembles were completely identical to the experimental ensemble, the Karplus equation parameters would result in a discrepancy between the calculated and experimentally measured scalar couplings. Because the C_α spin system in glycine differs profoundly from those of the other amino acids in having no attached C_β atom, it is plausible that coupling parameters involving the C_α atom that were developed for all residues would be the least applicable to glycine.

The difference between calculation and experiment for the ³J(C,C') coupling appears to be slightly more complicated, in that it is somewhat dependent on which parametrization is used. The experimentally measured value for the ³J(C,C') coupling is 0.26 Hz. For the orig. parametrization, the lowest possible value of the coupling is 0.44 Hz, while for the DFT1 and DFT2 parametrizations, the lowest possible values are 0.10 and 0.13 Hz, respectively. (For comparison, the maximum possible values range from 2.89 to 3.90 Hz.) Thus, matching the experimentally measured value would require the simulated ensemble to almost

Table 4. χ^2 Values for Calculated Scalar Couplings of Val₃ at 300 K for Various Force Field and Water Model Combinations^a

	all couplings			No ${}^3J(C,C')$ coupling		
	orig.	DFT1	DFT2	orig.	DFT1	DFT2
ff99SB (TIP3P)	2.00 (0.17)	2.99 (0.06)	2.31 (0.18)	1.22 (0.16)	2.50 (0.03)	1.64 (0.16)
ff99SB (TIP3P, RREMD)	1.88 (0.29)	3.13 (0.06)	2.18 (0.30)	1.11 (0.29)	2.67 (0.03)	1.54 (0.30)
ff99SB (TIP4P-Ew)	1.91 (0.25)	3.33 (0.17)	2.24 (0.29)	1.21 (0.29)	2.97 (0.20)	1.68 (0.33)
ff99SB (TIP4P-Ew, RREMD)	1.73 (0.15)	3.24 (0.09)	2.05 (0.16)	0.97 (0.18)	2.81 (0.08)	1.42 (0.18)
ff99SB* (TIP4P-Ew, RREMD)	1.69 (0.14)	3.09 (0.12)	1.98 (0.15)	0.92 (0.13)	2.64 (0.10)	1.33 (0.14)

^a Values are given as the means over four independent simulations, with the standard errors of the means given in parentheses. For RREMD simulations, probabilities are given as means over two independent simulations, with the differences between the two simulations given in parentheses.

exclusively sample ϕ angles of ± 60 – 90° , which is not commensurate with glycine's conformational flexibility. In addition, Graf et al. remark that there is often severe overlap in the spectra used to measure this coupling,³⁶ so it is possible that there is non-negligible error in the experimentally measured values themselves.

Given the above considerations, we recomputed the χ^2 values, this time excluding the ${}^2J(N',C_\alpha)$ and ${}^3J(C,C')$ couplings. We found that if these couplings were excluded, the χ^2 values dropped below 1.0 for the orig. and DFT2 calculated couplings and only slightly above 1.0 for the DFT1 calculated couplings (Table 3). Moreover, the differences between the various force field and solvent model combinations became relatively insignificant, although the TIP4P-Ew simulations still yielded better agreement with experimental results (Table 3). These data suggest that the ϕ and ψ dihedral angle potentials of the ff99SB force field are likely adequate for condensed phase simulation in TIP4P-Ew water.

Simulations of Val₃ at 300 K. Simulations of the hydrophobic Val₃ peptide probe a force field's ability to reproduce the conformational preferences of residues with side chains more complicated than the simple methyl group of alanine. This is not a trivial point, as backbone dihedral angle parameters are often developed using alanine di- or tetrapeptides,^{2,7,9,11,52} with the implicit assumption being that all residues with $C\beta$ atoms (with the exception of proline) will behave in essentially the same way and that any differences in the backbone preferences will be accounted for by additional interactions with the side chains.

As with Ala₃ and Gly₃, we performed four independent simulations of the Val₃ peptide using ff99SB and both water models. We found, however, that it was not possible to converge the simulations to our standard of less than 5% standard deviation in the mean interproton distances and the χ^2 values, even after 400 ns of simulation. Thus, we also performed RREMD simulations with these force field and solvent model combinations, as well as ff99SB* with TIP4P-Ew.

The results of these simulations suggest that the various force field and solvent model combinations produce conformational ensembles that are equivalently accurate within the statistical uncertainty, with no significant advantage for the TIP4P-Ew water model (Table 4). We do, however, find that the Karplus equation parameters developed using alanine peptides—particularly the DFT1 parameters—yield significantly larger χ^2 values (Table 4). These data suggest that backbone scalar coupling parameters derived from studies of only a single residue type may be inaccurate when applied to residues with different $C\beta$ configurations and that it may be preferable to use parameters that are “averaged” over all residue types (e.g., the orig. parameters) for assessing the accuracy of the conformational ensembles.

Because we encountered discrepancies with the ${}^3J(C,C')$ coupling in our studies of Gly₃ (see above), we examined the behavior of this coupling in our studies of Val₃. As can be seen in Table 4, it is clear that this coupling contributes a significant amount of error (reflected in the total χ^2 value), particularly for the orig. and DFT2 parameters. Moreover, if this coupling is excluded from the χ^2 calculation, then the simulated Val₃ ensembles generate χ^2 values approximately equal to 1.0 (within the uncertainty bounds), which suggests that the simulated ensembles are equivalent to the experimental ensemble within the error of the Karplus equation parameters (Table 4). Nonetheless, the considerable variances in the χ^2 values, both with and without the ${}^3J(C,C')$ coupling, make it difficult to conclude which force field and water model combination yields more accurate conformational ensembles for Val₃, despite significant differences in the conformational preferences of the central residue (Supporting Information Table 5).

Development of an Optimized ϕ' Backbone Dihedral Angle Potential. While the χ^2 values for Gly₃ indicate that the combination of the ff99SB force field and the TIP4P-Ew water model accurately simulates glycine residues (after excluding the problematic couplings from the calculation), the data we obtained for Ala₃ and, to a lesser extent, Val₃ clearly suggest that there is room for improving the intrinsic conformational preferences of nonglycine residues. More specifically, our Ala₃ data imply that an increase in sampling of the PPII conformation could yield better agreement with the experimental data. Similar observations were made previously by Wickstrom et al., who found that increased sampling of the PPII conformation in Ala₃ and Ala₅ in TIP4P-Ew water yielded improved agreement with the scalar coupling measurements and suggested that further increases in sampling the PPII conformation would be desirable.²⁷

There are several potential avenues for improving the intrinsic conformational preferences of amino acids given a specific force field and water model combination, including modifying partial atomic charges, van der Waals parameters, dihedral angle potentials, and other bonded interaction parameters. Of these, the two choices that would likely involve the least perturbation of the ff99SB force field's already excellent description of native state thermodynamics and dynamics (for folded proteins) would be the van der Waals parameters—as they apply to interactions with water molecules—and the backbone dihedral angle parameters.

In the course of our work, we derived optimized van der Waals (vdW) parameters for the interactions between TIP4P-Ew water molecules and alkane hydrogen and carbon atoms by fitting the vdW radii (R_i) and well depths (ϵ_i) of these atoms to bring the calculated solvation free energies of methane and *n*-butane into satisfactory agreement with experimentally determined values.

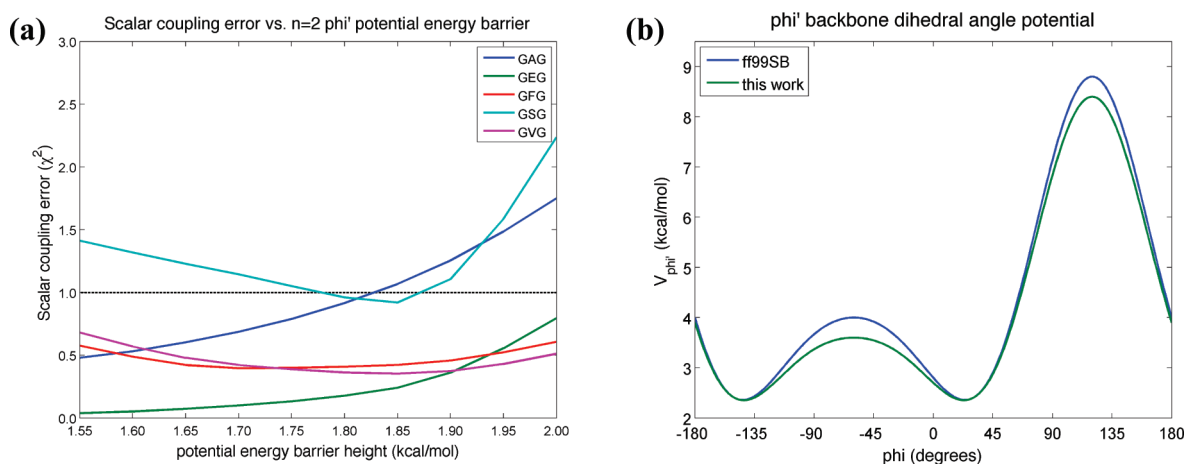


Figure 1. (a) Scalar coupling error (χ^2 value) as a function of the $n = 2$ term of the ϕ' backbone dihedral angle potential energy for GAG (dark blue), GEG (green), GFG (red), GSG (cyan), and GVG (magenta) peptides. (b) Potential energy as a function of ϕ for the ϕ' backbone potential, assuming a geometric relationship of $\phi = \phi' + 120^\circ$, for ff99SB (dark blue) and the optimized potential presented in this work (green).

Ultimately, we found that while these parameters yielded significantly more accurate solvation free energies for other alkane amino acid side chain analogues (e.g., propane and isobutane), they did not significantly improve the accuracy of the conformational ensemble for Val₃, the peptide that should benefit the most from such modifications due to its large alkane side chains (data not shown). We will further describe our methodology for developing optimized vdW parameters and present these data in future work.

The remaining possibility for improving intrinsic conformational preferences lies in the backbone dihedral angle potentials—the subject of numerous previous studies and force field development efforts.^{5–7,9,10} In the AMBER force fields, there are two sets of backbone dihedral angle potentials.⁷ One set of these potentials is based on the ϕ and ψ angles, defined as the torsions about the C–N–C α –C and N–C α –C–N atoms, respectively. These potentials apply to all residues in AMBER. The second set is based on the ϕ' and ψ' angles, defined as the torsions about the C–N–C α –C β and C β –C α –C–N atoms, respectively. These potentials apply to all *nonglycine* residues in AMBER. Given that we obtained χ^2 values less than 1 for Gly₃ (after excluding the problematic $^2J(N',C_\alpha)$ and $^3J(C,C')$ couplings), we elected to focus on improving the ϕ' and ψ' potentials rather than perturb the ϕ and ψ potentials.

There are two ways to modify these potentials that would increase sampling of the PPII conformation. First, one could shift the ψ' potential to increase sampling of both the β and PPII conformations, thereby decreasing the sampling of the α conformation. Recent studies, however, have suggested that the ff99SB force field may actually benefit from increased rather than decreased sampling of the α conformation.⁵ The second possibility is to shift the ϕ' potential to less frequently sample the β conformation and more frequently sample the PPII conformation. In principle, even large changes to the ϕ' potential would not greatly impact the overall sampling of the α conformation, although it would bias residues away from sampling conformations in the α basin with large negative ϕ values toward sampling those with ϕ angles near -60° . We therefore focused on optimizing the ϕ' potential to increase sampling of the PPII conformation.

Rather than introduce a new term to the ϕ' potential which would require determining an optimal energy magnitude and angular offset, we examined the existing ϕ' potential, which contains three terms, each with a different periodicity. The $n = 2$ term (i.e., the term with two maxima/minima over the range of ϕ') affects the height of the two potential energy maxima at $\phi = -60^\circ$ and 120° , assuming a geometric relationship of $\phi = \phi' + 120^\circ$. By modestly decreasing the magnitude of the $n = 2$ term, we could lower these barriers and increase the sampling of conformations near $\phi = -60^\circ$, thereby increasing the sampling of the PPII conformation. (Conformations at $\phi = 120^\circ$ are sufficiently high in energy due to both steric clash and the $n = 1$ term of the potential to preclude significant sampling.)

We performed REMD simulations of a number of different GXG peptides (where X = A, E, F, S, or V) in TIP4P-Ew water, varying the $n = 2$ potential energy term from 2.00 kcal/mol (the ff99SB value) down to 1.55 kcal/mol and calculating χ^2 values for predicted couplings. We used GXG peptides for the parametrization process because there are high quality coupling data available for them³⁷ and they provide a minimally perturbing context in which to examine the intrinsic conformational preferences of the central amino acids. In calculating the χ^2 values, we exclusively used the orig. Karplus equation parameters because they effectively average over all residue types, unlike the DFT1 and DFT2 parameters, which are optimized for alanine. We found that there was an improvement in the χ^2 values for all five peptides as the term was decreased to 1.85 kcal/mol and continued improvement for three of the five peptides (GAG, GEG, and GFG) to 1.80 kcal/mol (Figure 1a). Moreover, at 1.80 kcal/mol, the simulations of all five peptides yielded χ^2 values less than 1.0, indicating that the conformational preferences were accurate to within the limits of the Karplus equation parametrization (Figure 1a).

To validate this change to the $n = 2$ term of the ϕ' potential (Figure 1b), we performed REMD simulations using the aforementioned Val₃ peptide, as well as GLG and Ala₃ peptides, in TIP4P-Ew water. The Val₃ simulations indicate an overall improvement in the χ^2 value—larger than that observed for the optimized vdW parameters—but much of this improvement is in the predicted $^3J(C,C')$ coupling (Table 5). This result is somewhat unexpected in that the parametrization of the potential did not involve evaluation against any $^3J(C,C')$ couplings, as these couplings were not measured for the GXG peptides.³⁷

Table 5. χ^2 Values for Calculated Scalar Couplings of Val₃ at 300 K for Unmodified ff99SB and ff99SB with the Optimized ϕ' Backbone Dihedral Angle Potential, Both with TIP4P-Ew Water^a

	all couplings	no ³ J(C,C') coupling
ff99SB (unmodified)	1.73 (0.15)	0.97 (0.18)
opt. ϕ' dihedral potential	1.53 (0.02)	1.10 (0.00)

^a Values are given as the means over two independent RREMD simulations, with the differences between the two simulations given in parentheses. Only the orig. Karplus equation parameters are used for these calculations.

The GLG simulations display a modest improvement, with the χ^2 value (using only the orig. Karplus equation parameters) decreasing from 0.513 ± 0.004 to 0.483 ± 0.011 . This is concomitant with an increase in the probability of sampling the PPII conformation from 0.376 to 0.441, as well as a slight increase in sampling of α conformations (Supporting Information Table 6). The modified potential substantially improves the conformational ensembles for Ala₅, resulting in a 40–50% decrease in the χ^2 values computed with the orig. and DFT2 parameters (Table 6). This improvement is accompanied by the probability of sampling PPII conformation increase from 0.443 to 0.542, with no observed increase in the fraction of α conformations (Figure 2, Supporting Information Table 7).

Last, while this new parametrization is intended to improve intrinsic conformational preferences, it is important to know that it does not adversely affect native protein stability or dynamics. To verify this, we performed 60 ns simulations of ubiquitin, a well-characterized protein used in previous force field validation efforts. We observed no significant difference between our modified force field and ff99SB in examining either RMSDs from the crystal structure (Figure 3a) or the computed NMR S^2 order parameters (Figure 3b).

DISCUSSION

The majority of previous MD force field development and optimization efforts have focused on improving the agreement between gas phase *ab initio* and molecular mechanics calculations as the primary means of improving MD simulation accuracy.^{2,4,7–9,11,52} As simulations of proteins, nucleic acids, and other biomolecules are generally carried out in the condensed phase, however, one of the most pressing questions in the field is how accurately such parameters describe these molecules in the condensed phase—particularly in aqueous solution—and to what extent different solvent models may influence their structural ensembles.²⁵ Moreover, the accuracy of force fields and solvent models is paramount for simulations of intrinsically disordered proteins, as their manifold conformational states are similar in free energy²¹ and interactions with solvent are enhanced relative to folded proteins.²⁵

We therefore set out to assess the intrinsic conformational preferences of alanine, glycine, and valine in the AMBER ff99SB force field in combination with the TIP3P and TIP4P-Ew water models by simulating the Ala₃, Gly₃, and Val₃ peptides, respectively. In the cases of Ala₃ and Gly₃, the ff99SB/TIP4P-Ew combination yielded significantly more accurate conformational preferences than the ff99SB/TIP3P combination. For Val₃, the TIP4P-Ew simulations were systematically in better agreement with the experimental measurements, but the differences from

TIP3P were not statistically significant. Simulations of all three peptides demonstrated that the primary difference between the TIP3P and TIP4P-Ew ensembles is an increase in the extended (primarily PPII) conformations relative to more compact α -helical conformations.

In the case of Ala₃, the increase in sampling of the PPII conformation across the temperature range 275–350 K is unambiguously correlated with an improvement in agreement with NMR scalar coupling data. A similar correlation was noted in a previous study comparing the ensembles of Ala₃ and Ala₅ in TIP3P and TIP4P-Ew water at 300 K.²⁷ These observations are reinforced by data from another recent force field study, which demonstrated that the AMBER ff03/ff03* force fields yielded better agreement with scalar coupling data for the Ala₅ peptide than ff99SB/ff99SB* (in TIP3P water).⁵ One of the primary differences between the two force field families is the greater sampling of the PPII conformation by ff03/ff03* relative to ff99SB/ff99SB*.⁵ Moreover, these data are consistent with a multitude of experimental data regarding alanine in short peptides, which suggest that it primarily samples the PPII conformation.^{36,37,53–55}

While the Gly₃ results suggested that the intrinsic conformational preferences of glycine were already adequate using the ff99SB/TIP4P-Ew combination, the Ala₃ and Val₃ simulation suggested that further improvements were possible. We first explored the creation of van der Waals parameters optimized for simulation in TIP4P-Ew water but found that while they yielded significantly more accurate calculations of solvation free energy, they were unable to significantly improve the conformational ensemble of Val₃. We then considered a second approach—modifying the backbone dihedral angle potentials, specifically focusing on the ϕ' potential governing the balance between the β and PPII conformations. By lowering the energy scale of one term of this potential by 0.20 kcal/mol to increase sampling of the PPII conformation and performing REMD simulations of a variety of GXG peptides to assess the effects of our changes, we revised the ϕ' potential to yield more accurate intrinsic conformational preferences for a wide range of amino acids in TIP4P-Ew water while also maintaining the excellent native state stability of the ff99SB force field.

It is important to note that this revised potential is designed to increase the accuracy of the *intrinsic* conformational preferences of single amino acids, and there is no *a priori* reason that it should improve other types of conformational preferences, such as α -helix or β -sheet formation propensities, which are a combination of intrinsic conformational preferences and cooperative interactions between residues (e.g., hydrogen bonds). This makes our optimization strategy distinct from the ff99SB* modification of Best and Hummer, for example, which aims to improve the description of the helix-to-coil transition and uses fractional helicities of a longer α -helix-forming peptide in the parametrization process.⁵ As Best and Hummer have suggested, it is likely that additional physics and/or potentials must be introduced into current force fields to accurately capture such cooperative interactions.⁵

Another related issue that is often overlooked in the biomolecular simulation community is the use of solvent models with nonbonded interaction schemes that are different from those used in the parametrization of those solvent models. In particular, the TIP3P model was parametrized using simple truncation (cutoffs) for both electrostatic and van der Waals interactions,³⁰ whereas many current studies—including this work, as well as

Table 6. χ^2 Values for Calculated Scalar Couplings of Ala₅ at 300 K for Unmodified ff99SB and ff99SB with the Optimized ϕ' Backbone Dihedral Angle Potential, Both with TIP4P-Ew Water^a

	all couplings			no ³ J(C,C') coupling		
	orig.	DFT1	DFT2	orig.	DFT1	DFT2
ff99SB (unmodified)	2.44 (0.10)	1.87 (0.06)	2.14 (0.11)	1.73 (0.09)	1.20 (0.03)	1.37 (0.08)
opt. ϕ' dihedral potential	1.33 (0.04)	2.13 (0.05)	1.26 (0.02)	0.86 (0.05)	1.85 (0.05)	0.86 (0.02)

^a Values are given as the means over two independent RREMD simulations, with the differences between the two simulations given in parentheses.

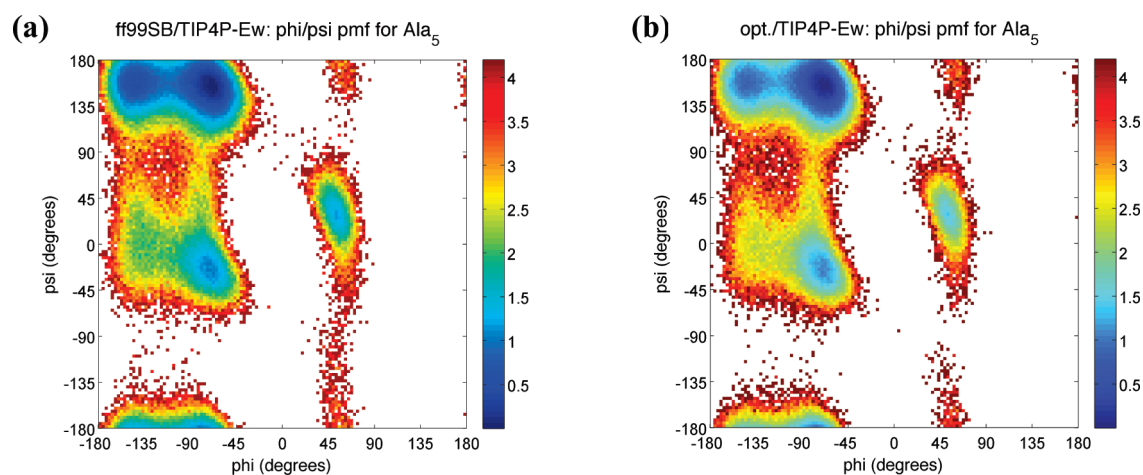


Figure 2. Average conformational preferences of residues 2–4 in the Ala₅ peptide using (a) the unmodified ff99SB force field or (b) the ff99SB force field with optimized ϕ' potential. Conformational preferences are represented as a potential of mean force (pmf), $W(\phi, \psi) = -RT \log p(\phi, \psi)$, with relative free energies given in kcal/mol.

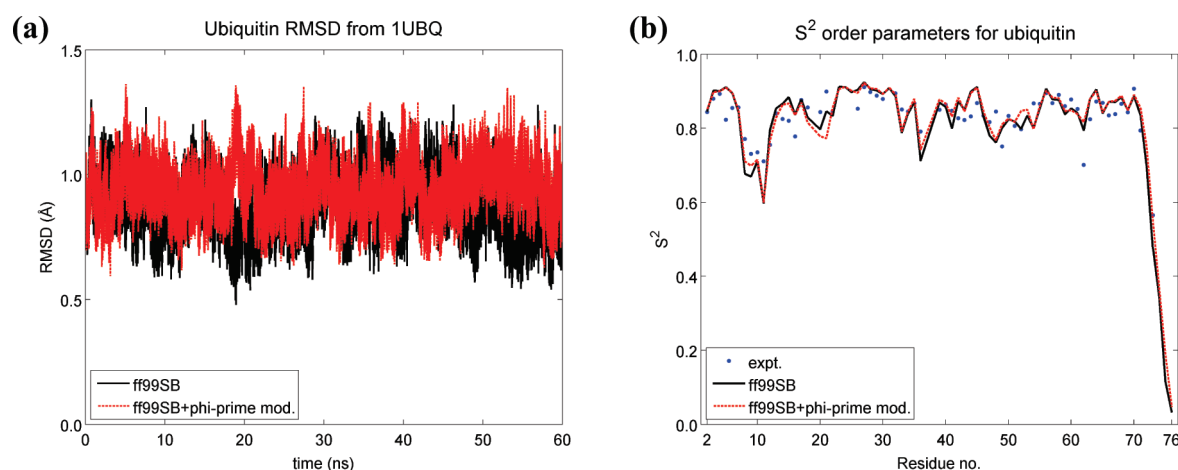


Figure 3. Native state stability and dynamics of unmodified ff99SB (black) and ff99SB with optimized ϕ' potential (red) for ubiquitin. (a) Root mean square distance (RMSD) to crystal structure 1UBQ⁴⁹ over 60 ns of simulation. The mean RMSDs are 0.88 Å for unmodified ff99SB and 0.94 Å for ff99SB with the optimized potential. (b) Lipari-Szabo order parameters (S^2) for ubiquitin at 300 K and pH 4.7, with experimentally derived isotropic values shown in blue dots.⁵⁰ The root-mean-square errors relative to the experimental S^2 values are 0.044 for both unmodified ff99SB and ff99SB with the optimized ϕ' potential.

other force field development and assessment studies^{5,7,27}—utilize TIP3P along with particle mesh Ewald (PME) for long-range electrostatics and corrections for van der Waals interactions beyond the direct-space cutoff. By contrast, TIP4P-Ew was derived specifically for use with modern simulation techniques, including PME and long-range van der Waals corrections.³³ It has been shown that the accuracy of the TIP3P model is

degraded under these simulation conditions,⁵⁶ and therefore it may be possible that the accuracy of the conformational ensembles generated with TIP3P could be improved if we were to revert to a simple truncated scheme for nonbonded interactions. Nonetheless, even under ideal simulation conditions, TIP3P does not reproduce experimentally measured characteristics of liquid water as accurately as TIP4P-Ew.^{33,56} Given the central role of

the solvent model in representing peptide–solvent interactions,²⁵ the optimization of current force fields for use with solvent models more advanced than TIP3P is logically sound and has already been suggested to be a promising avenue for force field development.⁶ More critically, while abandoning the use of PME would yield increased accuracy of the TIP3P model, it is known that disregarding long-range electrostatic interactions in MD simulations can lead to unphysical behavior of biomolecules.^{57–59} We therefore believe it both appropriate and sensible to have performed our assessments of TIP3P (and TIP4P-Ew) using PME and to have carried out our optimization efforts using TIP4P-Ew.

In addition to assessing the impact of solvent models on intrinsic conformational preferences, our simulations of Ala₃, Gly₃, and Val₃ revealed a number of insights into the calculation of experimental observables (scalar couplings) from structural ensembles. For Ala₃ (and to a lesser extent, Gly₃), we observed that the ff99SB/TIP4P-Ew combination yielded lower χ^2 values than the ff99SB/TIP3P *regardless* of the Karplus equation parameters used. For Val₃, however, decreases in the χ^2 values obtained with the orig. and DFT2 parameters were correlated with *increases* in χ^2 values obtained with the DFT1 parameters. While the orig. parameters were obtained by fitting measured NMR couplings to ϕ/ψ angles measured from high-resolution X-ray or NMR structures,^{60–63} the DFT1 and DFT2 parameters were derived from DFT calculations of the scalar couplings for the Ac–Ala–Nme dipeptide and NH₂–Ala–Ala–NH₂ peptide, respectively.⁶⁴ Thus, the orig. parameters implicitly average over all residue types, while the DFT1 and DFT2 parameters were derived explicitly using alanine residues. This in turn suggests that the orig. parameters may be applied with more or less equivalent accuracy to all amino acids, whereas the DFT1 and DFT2 parameters may be accurate for alanine residues, but their applicability to other residue types is in question. Moreover, the length of the peptide used for the Karplus equation parametrization may also matter. The modified ϕ' potential yielded improved χ^2 values for the Ala₅ peptide when computed using the orig. or DFT2 parameters, but higher χ^2 values when using the DFT1 parameters.

The need for residue-specific Karplus equation parameters is further exemplified by the case of Gly₃, in which the parameters for the ²J(N',C_α) coupling—obtained by fitting the Karplus equation with ϕ/ψ angles of a refined NMR structure⁶⁰—are not able to generate predicted couplings large enough to match the experimentally measured couplings. An examination of the original data for this set of Karplus equation parameters reveals that while there is some correlation between predicted and measured couplings, there is also a considerable spread between the results, especially for the residues in extended conformations.⁶⁰ We observed similar shortcomings using other residue-averaged parametrizations⁶³ for this coupling (data not shown).

Together, these data suggest that deriving residue-specific Karplus equation parameters would significantly improve the calculation of backbone couplings, particularly those involving C_α and/or C_β atoms, which have very different chemistries across the range of amino acids. It may be sufficient to derive a few sets of parameters (e.g., one set for the aromatic side chain residues, one set for the branched C_β side chain residues, etc.) rather than a unique set for each amino acid, but further study is needed to investigate the accuracy of such approaches. In addition, it is clear that the length of peptide used in the parametrization process affects the results and that dipeptides

may be of insufficient length to yield accurate parameters even for fairly short peptides (e.g., Ala₅). Calculating couplings with modestly larger peptides (e.g., GXG) may be sufficient to minimize the parametrization errors due to finite length. We intend to investigate both residue-specific and length effects in deriving Karplus equation parameters in future work.

CONCLUSIONS

Exciting new frontiers in biology, such as intrinsically disordered proteins, require an unprecedented interplay between simulation and experiment to fully understand the behavior of these biomolecules.^{35,65} This work and others demonstrate that current force field and water model combinations still require improvement to accurately describe disordered states^{26,27} and that such improvements may be realized by utilizing condensed phase simulations and experimental data to fine-tune parameters^{5,6} rather than relying solely on matching gas phase *ab initio* data, as has often been done in the past. A “hybrid” strategy of using gas phase *ab initio* data for initial parametrization and quantitative experimental data for fine-tuning those parameters may be valuable not only in the optimization of existing force fields but also in the development of next-generation fixed-charge and polarizable force fields. This force field development strategy, however, hinges upon a simultaneous development of more accurate methods for calculating experimental observables from simulated ensembles. Improvements in both of these areas will be critical for the interplay between simulation and experiment necessary for characterizing IDPs, as well as the overall advance of MD simulations as applied to biomolecules in general.

ASSOCIATED CONTENT

S Supporting Information. Calculated scalar couplings for Ala₃ at 275 K; conformational preferences of the central residues of Gly₃, Val₃, and Ala₅ at 300 K and of GLG at 298 K; conformational preferences of the first/N-terminal and third/C-terminal residues of Ala₃ at 275 K; and partial atomic charges for protonated C-terminal (–COOH) Ala, Gly, and Val residues. This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: tthead-gordon@lbl.gov.

ACKNOWLEDGMENT

The work reported here is supported by the NSF Cyber-Infrastructure Award 0344670, as well as the resources of UC Berkeley CITRIS and the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

REFERENCES

- (1) van Gunsteren, W. F.; Dolenc, J. *Biochem. Soc. Trans.* **2008**, *36*, 11–15.
- (2) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

- (3) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.
- (4) 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.
- (5) Best, R. B.; Hummer, G. *J. Phys. Chem. B* **2009**, *113*, 9004–9015.
- (6) Best, R. B.; Mittal, J. *J. Phys. Chem. B* **2010**, *114*, 14916–14923.
- (7) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins* **2006**, *65*, 712–725.
- (8) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. *Proteins* **2010**, *78*, 1950–1958.
- (9) MacKerell, A. D., Jr.; Feig, M.; Brooks, C. L., III. *J. Comput. Chem.* **2004**, *25*, 1400–1415.
- (10) Sorin, E. J.; Pande, V. S. *Biophys. J.* **2005**, *88*, 2472–2493.
- (11) 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.
- (12) Mobley, D. L.; Dumont, E.; Chodera, J. D.; Dill, K. A. *J. Phys. Chem. B* **2007**, *111*, 2242–2254.
- (13) Ponder, J. W.; Case, D. A. *Adv. Protein Chem.* **2003**, *66*, 27–85.
- (14) Daggett, V. *Chem. Rev.* **2006**, *106*, 1898–1916.
- (15) Karplus, M.; Kuriyan, J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6679–6685.
- (16) Scheraga, H. A.; Khalili, M.; Liwo, A. *Annu. Rev. Phys. Chem.* **2007**, *58*, 57–83.
- (17) Dyson, H. J.; Wright, P. E. *Nat. Rev. Mol. Cell. Biol.* **2005**, *6*, 197–208.
- (18) Fink, A. L. *Curr. Opin. Struct. Biol.* **2005**, *15*, 35–41.
- (19) Uversky, V. N. *Protein Sci.* **2002**, *11*, 739–756.
- (20) Uversky, V. N.; Oldfield, C. J.; Dunker, A. K. *Annu. Rev. Biophys.* **2008**, *37*, 215–246.
- (21) Turoverov, K. K.; Kuznetsova, I. M.; Uversky, V. N. *Prog. Biophys. Mol. Biol.* **2010**, *102*, 73–84.
- (22) Dill, K. A.; Ozkan, S. B.; Shell, M. S.; Weikl, T. R. *Annu. Rev. Biophys.* **2008**, *37*, 289–316.
- (23) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. *Science* **1991**, *254*, 1598–1603.
- (24) Mao, A. H.; Crick, S. L.; Vitalis, A.; Chicoine, C. L.; Pappu, R. V. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 8183–8188.
- (25) Florova, P.; Sklenovsky, P.; Banas, P.; Otyepka, M. *J. Chem. Theory Comput.* **2010**, *6*, 3569–3579.
- (26) Best, R. B.; Buchete, N. V.; Hummer, G. *Biophys. J.* **2008**, *95*, L07–L09.
- (27) Wickstrom, L.; Okur, A.; Simmerling, C. *Biophys. J.* **2009**, *97*, 853–856.
- (28) Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Shirts, M. R.; Dill, K. A. *J. Chem. Theory Comput.* **2009**, *5*, 350–358.
- (29) Shirts, M. R.; Pande, V. S. *J. Chem. Phys.* **2005**, *122*, 134508.
- (30) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (31) Abascal, J. L.; Vega, C. *J. Chem. Phys.* **2005**, *123*, 234505.
- (32) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269–6271.
- (33) Horn, H. W.; Swope, W. C.; Pitner, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. *J. Chem. Phys.* **2004**, *120*, 9665–9678.
- (34) Rick, S. W. *J. Chem. Phys.* **2004**, *120*, 6085–6093.
- (35) Fawzi, N. L.; Phillips, A. H.; Ruscio, J. Z.; Doucleff, M.; Wemmer, D. E.; Head-Gordon, T. *J. Am. Chem. Soc.* **2008**, *130*, 6145–6158.
- (36) Graf, J.; Nguyen, P. H.; Stock, G.; Schwalbe, H. *J. Am. Chem. Soc.* **2007**, *129*, 1179–1189.
- (37) Hagarman, A.; Measey, T. J.; Mathieu, D.; Schwalbe, H.; Schweitzer-Stenner, R. *J. Am. Chem. Soc.* **2010**, *132*, 540–551.
- (38) 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. J.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347–1363.
- (39) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Kollman, P. A. *J. Am. Chem. Soc.* **1993**, *115*, 9620–9631.
- (40) Vanquelef, E.; Simon, S.; Marquant, G.; Garcia, E.; Klimerek, G.; Delepine, J. C.; Cieplak, P.; Dupradeau, F.-Y. *R.E.D. Server: a web service designed to derive RESP and ESP charges and to generate force field libraries for new molecules/molecular fragments*; Université de Picardie Jules Verne: Amiens, France; Sanford-Burnham Institute for Medical Research: La Jolla, CA, 2010.
- (41) Case, D. A.; Darden, T. A.; Cheatham, I.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossvary, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M.-J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. *AMBER 11*; University of California: San Francisco, 2010.
- (42) Joung, I. S.; Cheatham, T. E., III. *J. Phys. Chem. B* **2008**, *112*, 9020–9041.
- (43) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.
- (44) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593.
- (45) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.
- (46) Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
- (47) Okur, A.; Roe, D. R.; Cui, G. L.; Hornak, V.; Simmerling, C. *J. Chem. Theory Comput.* **2007**, *3*, 557–568.
- (48) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.
- (49) Vijay-Kumar, S.; Bugg, C. E.; Cook, W. J. *J. Mol. Biol.* **1987**, *194*, 531–544.
- (50) Tjandra, N.; Feller, S. E.; Pastor, R. W.; Bax, A. *J. Am. Chem. Soc.* **1995**, *117*, 12562–12566.
- (51) Lipari, G.; Szabo, A. *J. Am. Chem. Soc.* **1982**, *104*, 4546–4559.
- (52) Wang, J. M.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049–1074.
- (53) Schweitzer-Stenner, R. *J. Phys. Chem. B* **2009**, *113*, 2922–2932.
- (54) Shi, Z.; Chen, K.; Liu, Z.; Ng, A.; Bracken, W. C.; Kallenbach, N. R. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 17964–17968.
- (55) Shi, Z.; Olson, C. A.; Rose, G. D.; Baldwin, R. L.; Kallenbach, N. R. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 9190–9195.
- (56) Price, D. J.; Brooks, C. L., III. *J. Chem. Phys.* **2004**, *121*, 10096–10103.
- (57) Cheatham, T. E.; Miller, J. L.; Fox, T.; Darden, T. A.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 4193–4194.
- (58) Norberg, J.; Nilsson, L. *Biophys. J.* **2000**, *79*, 1537–1553.
- (59) Zuegg, J.; Gready, J. E. *Biochemistry* **1999**, *38*, 13862–13876.
- (60) Ding, K.; Gronenborn, A. M. *J. Am. Chem. Soc.* **2004**, *126*, 6232–6233.
- (61) Hennig, M.; Bermel, W.; Schwalbe, H.; Griesinger, C. *J. Am. Chem. Soc.* **2000**, *122*, 6268–6277.
- (62) Hu, J. S.; Bax, A. *J. Am. Chem. Soc.* **1997**, *119*, 6360–6368.
- (63) Wimer, J.; Schwalbe, H. *J. Biomol. NMR* **2002**, *23*, 47–55.
- (64) Case, D. A.; Scheurer, C.; Bruschweiler, R. *J. Am. Chem. Soc.* **2000**, *122*, 10390–10397.
- (65) Sgourakis, N. G.; Merced-Serrano, M.; Boutsidis, C.; Drineas, P.; Du, Z.; Wang, C.; Garcia, A. E. *J. Mol. Biol.* **2011**, *405*, 570–583.

Electron Localization Function at the Correlated Level: A Natural Orbital Formulation [*Journal of Chemical Theory and Computation* **2010**, *6*, 2736–2742 DOI: 10.1021/ct1003548].

Ferran Feixas, Eduard Matito,* Miquel Duran, Miquel Solà, and Bernard Silvi*

In a paper recently published,¹ we stated that Piris had developed a functional which did not fulfill either the sum rules or the antisymmetry requirements. The formula we gave was wrong. The actual formula we should have written is

$$\begin{aligned} \pi(\mathbf{r}_1, \mathbf{r}_2) = & \sum_i \sum_{j,i} n_i n_j \varphi_i^*(\mathbf{r}_1) \varphi_j^*(\mathbf{r}_2) \varphi_i(\mathbf{r}_1) \varphi_j(\mathbf{r}_2) \\ & - \sum_i \sum_{j \neq i} \sqrt{n_i n_j} \varphi_i^*(\mathbf{r}_1) \varphi_j^*(\mathbf{r}_2) \varphi_j(\mathbf{r}_1) \varphi_i(\mathbf{r}_2) \\ & - \sum_i \sum_{j \neq i}^{\text{nco}} \sqrt{(1-n_i)(1-n_j)} \varphi_i^*(\mathbf{r}_1) \varphi_j^*(\mathbf{r}_2) \varphi_j(\mathbf{r}_1) \varphi_i(\mathbf{r}_2) \\ & + 2 \sum_{i,j > \text{nco}} \sum_{j \neq i} \sqrt{n_i n_j} \varphi_i^*(\mathbf{r}_1) \varphi_j^*(\mathbf{r}_2) \varphi_j(\mathbf{r}_1) \varphi_i(\mathbf{r}_2) \\ & - \frac{1}{2} \sum_i n_i^2 \varphi_i^*(\mathbf{r}_1) \varphi_i^*(\mathbf{r}_2) \varphi_i(\mathbf{r}_1) \varphi_i(\mathbf{r}_2) \end{aligned}$$

where nco is the number of occupied HF orbitals, and which does not fulfill the sum rule. Notice this is not the expression of PNOF1,² which attains both the sum rule and the antisymmetry prescriptions, but a particular application of PNOF1 that one can deduce from eqs 61, 62, and 40 in ref 2. PNOF1 depends upon a matrix (Δ) which is not specified and thus cannot be used, except for the particular application we just mentioned. However, this one is not convenient for the calculation of the electron localization function (ELF) because it fulfills neither the sum rule nor provides a correct description of the Fermi hole. The recently developed PNOF2,³ PNOF3,⁴ and PNOF4⁵ functionals also fulfill the aforementioned requirements, and in principle, they could be used for the computation of the pair density needed in the calculation of the ELF.

This functional was not used for the calculations presented in the paper, and thus the conclusions drawn in the paper hold. We apologize for this mistake and thank Prof. Mario Piris for noticing this error.

REFERENCES

- (1) Feixas, F.; Matito, E.; Duran, M.; Solà, M.; Silvi, B. *J. Chem. Theory Comput.* **2010**, *6*, 2736–2742.
- (2) Piris, M. *Int. J. Quantum Chem.* **2006**, *106*, 1093–1104.
- (3) Piris, M.; Lopez, X.; Ugalde, J. *J. Chem. Phys.* **2007**, *126*, 214103.
- (4) Piris, M.; Matxain, J.; Lopez, X.; Ugalde, J. *J. Chem. Phys.* **2010**, *132*, 031103.
- (5) Piris, M.; Matxain, J.; Lopez, X.; Ugalde, J. *J. Chem. Phys.* **2010**, *133*, 111101.

DOI: 10.1021/ct2001123

Published on Web 03/04/2011